STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 897–898

FROM THE EDITOR

This issue is devoted mainly, first to papers on *survey sampling* prepared by the authors participating in the ***Baltic-Nordic Conference on Survey Sampling*** in Ammarnäs, a small village in the north of Sweden in August, 2002. The rest of the papers will be published in the June 2003 issue of our journal. This part of the issue, **Survey Sampling (1),** was organised and edited by Daniel Thorburn, Dep. of Statistics, University of Stockholm. He has also prepared the *Foreword,* which is presented after my comments upon this issue.

There are also three articles in section ***Other Articles*** in this issue, one ***book review***, information on *economic and social trends in transition countries, obituary* and *acknowledgements of referees of volume 5***.**

There are following seven articles devoted to some problems of sampling survey:

1. M. Carlson, *Assessing Microdata Disclosure Risk Using the Poisson-Inverse Gaussian Distribution.*
2. M. Davidsen and M. Kjøller,. *The Danish Health and Morbidity Survey 2000 – Design and Analysis.*
3. D. Hedlin*, Estimating Totals in some U.K. Business Surveys.*
4. A. Holmberg, *A Multiparameter Perspective on the Choice of Sampling Design in Surveys.*
5. V. Kiviniemi and R. Lehtonen, *Web Tools in Teaching and Learning of Survey Sampling: The VliSS Application.*
6. S. Laaksonen, *Traditional and New Techniques for Imputation.*
7. K. Meister, *Asymptotic Considerations Concerning Real Time Sampling Methods.*

The second part of this issue under the title **Other articles** contains three articles also devoted to sampling survey:

1. B. Prasad, R.S. Singh and H. P. Singh, *Modified Chain Ratio Estimators for Finite Population Mean Using Two Auxiliary Variables in Double Sampling*
2. G.N. Singh, *Estimation of Population Ratio in Two Phase Sampling.*
3. T.P. Tripathi, H. P. Singh and L. N. Upadhyaya, *A General Method of Estimation and its Application to the Estimation of Coefficient of Variation*.

The part **Book Reviews** contains a review of: W. Charemza and K Strzala (Eds), ***East European Transition and EU Enlargement: A Quantitative***

Approach, (prepared by Subrata Ghatak). The book contains a collection of papers presented at the seminar held in July 2000, in Gdansk, Poland. Our journal announced this seminar in vol. 4, Number 6, December 2000.

Next part of this issue is devoted to **Economic and Social Trends in Transition Countries**. This part deals with *Basic Economic Trends in Countries of Central and Eastern Europe and CIS Countries* prepared by M. Bieńkowska, E. Czumaj and J. Gniadzik from the Central Statistical Office of Poland.

The editor announces with the deepest sorrow that Professor Tore L. Dalenius passed away in January 2002. He remembers his first meeting with Professor Dalenius in Vienna in 1973. **Obituary** of Professor Tore Dalenius is given in this issue.

The issue is concluded with the *Acknowledgements of referees* of Volume 5.

Jan Kordos
The Editor

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 899

SURVEY SAMPLING (1)

Foreword

In the summer of 2002 we arranged a ***Baltic-Nordic Conference on Survey Sampling*** in Ammarnäs, a small village in the north of Sweden. Many statisticians from all the seven countries of this region participated and many of them presented interesting papers. I do not think that any participant regretted participating. For those who were not so lucky, we have invited some of the best presentations to this special issue of *Statistics in Transition* and I feel confident that many readers will benefit from them.

The papers show the diversity and depth of the interest in survey sampling in this part of the world. There are applications from areas like health surveys (Davidssen & Kjöller) and business surveys (Hedlin). There are papers devoted to special aspects of survey sampling such as data security (Carlson) and imputation (Laaksonen) as well as papers on more mainstream theoretical sampling methods (Holmberg and Meister). Finally there is one paper on how to teach survey sampling using Internet. The papers also show the diversity by coming from six different countries and within the countries from both universities and statistical agencies. I hope that all readers will benefit from reading the contributions. Together they show that you cannot make good applications without a thorough theoretical knowledge and you cannot make good theory without a profound knowledge of statistical practice. All papers are certainly relevant for transition countries but also for survey statisticians from other countries. There are still some papers from the conference in the editorial process and I hope that they will be published in a later issue.

Finally, I want to thank all the participants to the conference for making it so interesting and *Statistics in Transition* for publishing this issue.

Daniel Thorburn
Dep of Statistics University of Stockholm
S-106 91 Stockholm, Sweden

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 901–925

ASSESSING MICRODATA DISCLOSURE RISK USING THE POISSON-INVERSE GAUSSIAN DISTRIBUTION

Michael Carlson[1]

ABSTRACT

An important measure of identification risk associated with the release of microdata or large complex tables is the number or proportion of population units that can be uniquely identified by some set of characterizing attributes which partition the population into subpopulations or cells. Various methods for estimating this quantity based on sample data have been proposed in the literature by means of superpopulation models. In the present paper the Poisson-inverse Gaussian (PiG) distribution is proposed as a possible approach within this context. Disclosure risk measures are discussed and derived under the proposed model as are various methods of estimation. An example on real data is given and the results indicate that the PiG model may be a useful alternative to other models.

Key words: statistical disclosure; uniqueness; inverse-Gaussian; Poisson-mixture; superpopulation.

1. Introduction

A considerable amount of research has been done in the area of statistical disclosure and different approaches to defining and assessing disclosure risk are treated in depth by among others Dalenius (1977), Duncan and Pearson (1991), Frank (1976,1988), Lambert (1993), Skinner et al. (1994), Willenborg and de Waal, (1996, 2000). Recent publications include Doyle et al. (2001) and Domingo-Ferrer (2002). A special case concerns the release of public-use microdata files and so-called identity disclosure, see Duncan and Lambert (1989). It is well known that there remains the possibility that statistical disclosure could occur even when such a file has been anonymized by the removal of direct identifiers, see e.g. Block and Olsson (1976) and Dalenius (1986), although Blien

[1] Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden.
e-mail: Michael.Carlson@stat.su.se

et al. (1993) demonstrated that it may be difficult in practice. The main concern is to ensure that no record in a released microdata set can be reliably associated with an identifiable individual.

For instance, a unique is defined as an entity that has a unique set of values on a set of characterizing attributes or key attributes. A unit that is unique in the population is referred to as a population unique whereas a unit that is unique in a sample is referred to as a sample unique. If a population unique is included in the sample it is necessarily also a sample unique but the converse does not hold. Obviously a population unique is subjected to a greater risk of being exposed relative to other non-unique units if included in the released data. Furthermore, it could be argued that an intruder will be more inclined to focus upon those records that are sample unique since it is only these that can by definition be population uniques, see e.g. Skinner et al. (1994) and Elliot et al. (1998). Thus, a possible indicator of the identification risk associated with the release of microdata is the number or proportion of population uniques included in the sample amongst the sample uniques.

The objective is to estimate this proportion based on sample information, e.g. a data set considered for release. Various methods for estimating this quantity based on sample data have been proposed in the literature by means of superpopulation models and especially compound Poisson models. Under a superpopulation model it is assumed that the population at hand, as defined by the frequency structure of the key attributes, has been generated by some appropriate distribution. The risk assessment, here in terms of uniqueness, is then reduced to a matter of parameter estimation and prediction. Bethlehem et al. (1990) were perhaps the first to adapt a superpopulation approach and others include Chen and Keller-McNulty (1998), Hoshino (2001), Samuels (1998), Skinner and Holmes (1993, 1998), St-Cyr (1998) and Takemura (1999).

In the present paper we propose the Poisson-inverse Gaussian (PiG) distribution as a possible candidate. This distribution has appeared elsewhere in the literature but we are not aware of it being applied to the disclosure problem earlier. It was introduced by Holla (1966) in studies of repeated accidents and recurrent disease symptoms. Sichel (1971) developed the PiG to a more general three-parameter family of distributions and applied it to density and size distributions of diamonds, sentence-length and word frequency data and to model repeat-buying behavior, (Sichel, 1973, 1974, 1975, and 1982a). Ord and Whitmore (1986) evaluated the PiG as an alternative to other distributions for species abundance data and Willmot (1987) for modelling insurance claim data. Chen and Keller-McNulty (1998) noted that, in practice, the frequency distribution in disclosure applications tends to have an inverse J-shape with heavy upper tail. St-Cyr (1998) also describes this typical behavior. Since the PiG distribution is characterized by its positive skewness and heavy upper tail it appears to be an appropriate distribution for modeling frequency counts in disclosure applications. Furthermore, the PiG distribution is expressed in closed

form which gives it an advantage over e.g. the lognormal which requires numerical integration.

Here we will limit the scope to a theoretical discussion of the PiG model with a simple example on real data to illustrate the method. An evaluation of the model with real-life data examples and its competitiveness with alternative approaches is intended to appear in a separate report. In the following section some basic notation is introduced and the superpopulation model is specified. In section 3 the PiG distribution is reviewed and in section 4 its application to the problem of assessing disclosure risks, here in terms of uniqueness, is discussed. Parameter estimation is described in section 5 and in section 6 the results of the empirical example are reported. Some concluding remarks and directions for future research are given in section 7.

2. Specification of the Superpopulation Model

2.1. Basic notation

Consider a finite population U of size N from which a simple random sample $s \subset U$ of size $n \leq N$ is drawn. The sampling fraction is denoted by $\pi_s = n/N$. With each unit $h \in U$ is associated the values of a number of discrete variables, $Z_1,\ldots,Z_q$ with $C_1,\ldots,C_q$ categories respectively. The cross-classification of these variables define the discrete variable X with $\prod C_i = C$ categories or cells and for simplicity we let the cells of X be labeled as $1,\ldots,C$.

Following e.g. Bethlehem et al. (1990) the Z_i are termed key variables, X the key and the C different categories of X, the key values. Thus, the key divides the population into C subpopulations $U_i \subseteq U$ and by F_i we denote the number of units belonging to subpopulation U_i, i.e. the population frequency or size of cell i The sample counterpart is analogously defined and denoted by f_i.

Define T_j and its sample counterpart t_j as the number of cells of size j, i.e.

$$T_j = \sum_{i=1}^{C} I_{F_i = j} = \#(i ; F_i = j), \quad j = 0, 1,\ldots, N$$

and

$$t_j = \sum_{i=1}^{C} I_{f_i = j} = \#(i ; f_i = j), \quad j = 0, 1,\ldots, n$$

respectively and where $I_{(\bullet)}$ denotes the usual indicator function. The T_j and t_j are usually termed cell size indices or frequencies of frequencies and correspond to the equivalence classes of Greenberg and Zayatz (1992). It is clear that

$$\sum_{i=1}^{C} F_i = \sum_{i=1}^{N} jT_j = N, \qquad \sum_{i=1}^{C} f_i = \sum_{i=1}^{n} jt_j = N$$

and

$$\sum_{i=0}^{N} T_j = \sum_{i=0}^{n} t_j = C.$$

Of these quantities, C, N and n are fixed in the design, the f_i and t_j are observed and the F_i and T_j are assumed to be unknown. The goal is to model and estimate the population frequency structure, i.e. the T_j and especially T_1 which is the number of unique individuals in the population, based on sample information.

2.2. Superpopulation model

The frequency structure, $\{T_j\}$, is a function of the actual F_i which are unknown and therefore need to be estimated from the observed f_i and t_j. However, as St-Cyr (1998) commented, it is not possible for a sample to carry all the information about the structure of a population and finite population theory will give only unreliable estimates when the sampling fraction is small. See also the discussion by Willenborg and de Waal (2000, p. 55). A way out is to model the frequency structure and view the population as the realization of a superpopulation model.

As a starting point we therefore assume that the cell frequencies are generated independently from Poisson distributions with individual rates λ_i, $i = 1,\ldots,C$. The Poisson model is motivated by thinking of the N units in the population as falling into the C different cells with probability of the ith cell denoted by π_i. Given the N, C and the π_i the frequencies will follow a multinomial distribution and if the number of cells is large enough each cell frequency is approximately independently binomial with parameters N and π_i. Since the population size is usually quite large and the π_i small due to large C the Poisson distribution is used to approximate the binomial with $\lambda_i = N\pi_i$.

To simplify the model further we view the λ_i as independent realizations of a continuous random variable Λ with a common probability density function (pdf) $g(\lambda)$. The number of cells C is usually quite large and this assumption will significantly reduce the number of parameters that need to be estimated. The simplification seems reasonable also in view of that we are not conditioning on any of the characterizing variables defining the key.

The specification of the mixing distribution $g(\lambda)$ is the crucial step and several different suggestions have been studied. Bethlehem et al. (1990) proposed a gamma distribution which implies that the marginal distribution of each F_i is the negative-binomial. This model has however been noted to provide a poor fit to real-life data. Skinner and Holmes (1993, 1998) argue instead for the use of a lognormal distribution. Chen and Keller-McNulty (1998) considered a shifted negative-binomial for the marginal distribution of the F_i and achieve better results compared to the Poisson-gamma. St-Cyr (1998) proposed a mixture of a Pareto and a truncated lognormal distribution based on his findings from studying the

relationship between the conditional probability of population uniqueness and the sampling fraction. Hoshino and Takemura (1998) investigated the relationships between different models and provide interesting results.

It is also obvious in many situations that certain combinations of the key variables may be impossible, such as married 4-year-olds or male primiparas, i.e. so-called structural zeroes. Skinner and Holmes (1993) specified their superpopulation model to allow for individual rates to be zero with positive probability. Following their idea we therefore assume that the distribution of the λ_i is a mixture θ of a discrete probability mass at zero $(1-\theta)$ of a continuous distribution with pdf $g(\lambda)$ on $(0,\infty)$, i.e. we specify the marginal of the cell frequencies as

$$\Pr(F_i = j) \;=\; \theta I_{j=0} + (1-\theta)P_j \tag{1}$$

where $0 \le \theta < 1$ and

$$P_j = \int_0^\infty \frac{\lambda^j e^{-\lambda}}{j!}\, g(\lambda) d\lambda \,. \tag{2}$$

Note that with the Poisson assumption the total of the cell counts ΣF_i is a random variable. Although it is true from the design that $\Sigma F_i = N$ it may suffice to check if $\Sigma \mathrm{E}(F_i) = N$, rather than conditioning on the actual population size, cf. Bethlehem et al. (1990). Denoting the expectation of the distribution in (2) by μ this requirement translates to

$$\mathrm{C}(1-\theta)\mu = N. \tag{3}$$

3. The Poisson-Inverse Gaussian Distribution

We assume that the individual rates follow an inverse Gaussian (iG) distribution and using the same parameterization as Willmot (1987) the pdf is

$$g(\lambda \mid \mu, \tau) = \frac{\mu}{(2\pi\tau\lambda^3)^{1/2}} \exp\left(-\frac{(\lambda-\mu)^2}{2\tau\lambda} \right), \quad \lambda > 0 \tag{4}$$

where $\mu > 0$, $\tau > 0$. The mean and variance of the iG are μ and $\mu\tau$ respectively. Folks and Chhikara (1978) provide a review of the iG distribution with an extensive set of references. See also Johnson et al. (1994, chapter 15). The inverse Gaussian distribution appears as e.g. the first passage time distribution of Brownian motion with positive drift. It may also be derived through an inversion relationship associated with cumulant generating functions of the Gaussian and inverse Gaussian families, hence the name.

Integrating out λ from (2) with $g(\lambda)$ replaced by (4) yields the probability mass function (pmf) of the Poisson-inverse Gaussian (PiG), i.e.

$$P_j = \frac{\sqrt{\eta}}{j!}\left(\frac{\mu}{\eta}\right)^j \frac{\mathrm{K}_{j-1/2}(\mu\eta/\tau)}{\mathrm{K}_{-1/2}(\mu/\tau)}, \quad j = 0,1,\ldots \tag{5}$$

where $\eta = (1+2\tau)^{1/2}$ and $\mathrm{K}_\gamma(z)$ denotes a modified Bessel function of the third kind (sometimes referred to as the second kind) of order γ and with argument z, see Abramowitz and Stegun, (1970, chapter 10). The mean and variance of the PiG distribution are μ and $\mu(1+\tau)$ respectively.

The expression in (5) is not always practical due to the Bessel functions, but by using that $\mathrm{K}_{-1/2}(z) = \mathrm{K}_{1/2}(z) = (\pi/2z)^{1/2}\exp(-z)$, we note that the first two probabilities are

$$P_0 = \exp\left(\frac{\mu}{\tau}(1-\eta)\right) \quad \text{and} \quad P_1 = \frac{\mu}{\eta}P_0. \tag{6}$$

Using also that $\mathrm{K}_\gamma(z) = \mathrm{K}_\gamma(z)$ and the recurrence relationship $\mathrm{K}_{\gamma+1}(z) = (2\gamma/z)\,\mathrm{K}_\gamma(z) + \mathrm{K}_{\gamma-1}(z)$, a more practical recurrence formula for calculating the probabilities is given by

$$P_j = \frac{\tau}{\eta^2}\frac{2j-3}{j}P_{j-1} + \frac{\mu^2}{\eta^2}\frac{1}{j(j-1)}P_{j-2}, \quad j = 2,3,\ldots. \tag{7}$$

3.1. Sampling From the Population

Assume that the population level cell frequencies are generated from a PiG model. Under simple random sampling without replacement, the sampling distribution of the cell frequencies of the PiG is hard to manipulate. We therefore assume Bernoulli sampling, cf. e.g. Särndal et al. (1992, chapter 3), in which each unit is drawn independently from the population with equal $\pi_s = n/N$ as a convenient approximation. This yields

$$f_i \mid \lambda_i \sim \mathrm{Po}(\pi_s\lambda_i) \quad \text{and} \quad F_i - f_i \mid \lambda_i \sim \mathrm{Po}((1-\pi_s)\lambda_i)$$

and

$$f_i \mid F_i \sim \mathrm{Bin}(F_i,\pi_s). \tag{8}$$

It is then easily seen that the marginal pmf for the sample cell frequencies f_i is defined by

$$\Pr(f_i = j) = \theta I_{j=0} + (1+\theta)p_j \tag{9}$$

where

$$p_j = \int_0^\infty \frac{(\pi_s\lambda)^j e^{-\pi_s\lambda}}{j!}\frac{\mu}{\left(2\pi\tau\lambda^3\right)^{1/2}}\exp\left(-\frac{(\lambda-\mu)^2}{2\tau\lambda}\right)d\lambda$$

$$= \int_0^\infty \frac{\lambda^j e^{-\lambda}}{j!} \frac{\mu_s}{\left(2\pi\tau_s\lambda^3\right)^{1/2}} \exp\left(-\frac{(\lambda-\mu_s)^2}{2\tau_s\lambda}\right) d\lambda \qquad (10)$$

i.e. a PiG distribution where $\mu_s = \pi_s\mu$, $\tau_s = \pi_s\tau$ and defining $\eta_s = (1+2\tau_s)^{1/2}$; the second line in (10) is derived by simple variable substitution. This provides an easy transformation when we have a sample from the larger population, i.e. given the sample we estimate the parameters μ_s and τ_s and simply multiply by π_s^{-1}. See section 2 of Sichel (1982a) and the discussion concerning sampling in Takemura (1999).

4. Risk assessment

The outline of the disclosure problem considered here is the same as that of many authors, e.g. Bethlehem et al. (1990), Elliot et al. (1998), Paass (1988) and Skinner and Holmes (1998). Consider an intruder who attempts to disclose information about a set of identifiable units in the population termed targets. The intruder is assumed to have prior information about the key values of the targets and attempts to establish a link between these and individual records in the released microdata file using the values of the key attributes. Assume that the intruder finds that a specific record r in the microdata file matches a target with respect to the key X. Now F_i is the number of units belonging to subpopulation U_i and we let $i(r)$ denote the value of X for record r. If $F_{i(r)}$ was known the intruder could infer that the probability of a correct link is $F_{i(r)}^{-1}$ and if $F_{i(r)} = 1$ the link is correct with absolute certainty.

Usually the intruder will not know the true value of $F_{i(r)}$ since the microdata set contains only a sample but by introducing a superpopulation model he may attach a probability distribution $\Pr(F_i = j)$ to the cell frequencies. Furthermore, it could be argued that an intruder will be more inclined to focus upon those records that are sample unique since it is only these that can by definition be population uniques. So equating disclosure risk with uniqueness, a simple measure of the risk for a given sample is the proportion of sample uniques that are also population uniques, i.e.

$$R = \frac{\#\left(\text{records that are population uniques } \textit{and} \text{ sample uniques}\right)}{\#\left(\text{records that are sample uniques}\right)}. \qquad (11)$$

Under simple random sampling or Bernoulli sampling the expected number of population uniques to fall into the sample is $\pi_s T_1 = nT_1/N$. Since T_1 is assumed unknown, the proportion (11) will have to be estimated. Under the model (1) the expected number of population uniques is

$$\mathrm{E}(T_1) \;=\; C\,\Pr(F_i = 1) \;=\; \mathrm{C}\,(1-\theta)P_1.$$

Thus, an obvious risk measure denoted by R1, is defined as the proportion of the observed number of sample uniques expected to be population uniques, i.e.

$$R_1 = \frac{\mathrm{E}(T_1)/N}{t_1/n} = \frac{\pi_s C(1-\theta)}{t_1}\frac{\mu}{\eta}\exp\left(\frac{\mu}{\tau}(1-\eta)\right). \tag{12}$$

where we have used (6). Again assuming Bernoulli sampling, an alternative risk measure R_2 follows naturally from the conditional pmf of F_i given f_i, i.e. from (5), (8) and (10) we have

$$R_2 = \Pr(F_i = 1 \mid f_i = 1) = \frac{\pi_s \Pr(F_i = 1)}{\Pr(f_i = 1)} = \frac{\mathrm{E}(T_1)/N}{\mathrm{E}(t_1)/n} \tag{13}$$

which simplifies to the risk measure

$$R_2 = \frac{\eta_s}{\eta}\exp\left(\frac{\mu}{\tau}(\eta_s - \eta)\right) \tag{14}$$

i.e. the observed value of t_1 in (12) is replaced for its expectation under the model. This is the approach discussed by Skinner and Holmes (1998). Note that $R_2 \to 1$ as $\pi_s \to 1$ which is not necessarily the case with R_1. The risk measures R_1 and R_2 coincide if and only if a perfect fit of the model to the observed number of sample uniques is obtained, i.e. $\mathrm{E}(t_1) = t_1$. Either choice, the disclosure risk is estimated by replacing the parameters by their estimates into (12) or (14).

4.1. Extended Risk Measures

It should be noted that population uniqueness is neither a sufficient nor necessary condition for re-identification or for disclosing additional information, see e.g. Frank (1976) and Willenborg and de Waal (1996, pp. 19-20). It is not a sufficient condition since, first of all, the unique unit must be included in the sample and secondly, it must also be known to the intruder that the unit is in fact unique. It is not necessary for several reasons. If for instance a person in the population shares the same values on the key attributes with say only one other person, they will both be able to re-identify and disclose information about each other. In general, coalitions of respondents exchanging information can be formed within small groups sharing the same scores on the key attributes, in order to disclose information about an individual within the same group but outside the coalition. Alternatively, if a group of people share the same values on the key attributes, none of them are unique. But if they in addition all share the same score on a certain sensitive attribute provided in the released data, the sensitive information can be disclosed for all the individuals in that group without re-identifying individual records. Another possibility is response knowledge, i.e. knowledge that a specific individual participated in the survey and consequently that his or her data must be included in the data. Identification and disclosure can

then occur if the person is unique in the sample and not necessarily in the population (Bethlehem et al., 1990).

These issues will not be investigated further in the present paper, but they do however motivate an extension of the risk measure in (13) to a more general measure defined by

$$\Pr(F_i = j + k \mid f_i = j) \;=\; \frac{1}{k!}\frac{\int_0^\infty (1-\pi_s)\lambda^{k+j}\exp(-\lambda)g(\lambda)d\lambda}{\int_0^\infty \lambda^{j}\exp(-\pi_s\lambda)g(\lambda)d\lambda}$$

for j = 1,2,… and k = 0,1,… . Simple examples pertaining to some of the situations just described may be e.g. j = 1 and k = 1 which after simplifying yields

$$\Pr(F_i = 2 \mid f_i = 1) \;=\; (1-\pi_s)\frac{(\mu\eta+\tau)}{\eta^2}R_2$$

or j = 2 and k = 0 which yields

$$\Pr(F_i = 2 \mid f_i = 2) \;=\; \pi_s\frac{\eta_s^2(\mu\eta+\tau)}{\eta^2(\mu_s\eta_s+\tau_s)}R_2.$$

where R_2 is the risk measure defined in (14).

5. Estimation

5.1. Moment Based Estimators

Let S_0 denote the number of structural zeroes. If this number is known a priori the parameter θ is known and equals S_0/C. When this is the case and especially when $\theta = 0$, the parameters μ_s and τ_s can be estimated using simple moment based approaches. Simple moment estimators are given by the sample mean and variance, i.e.

$$\tilde{\mu}_s = \frac{n}{C - S_0} \tag{15}$$

and

$$\tilde{\tau}_{s,1} = \frac{1}{\tilde{\mu}_s(C-S_0)}\sum_{j=0}^{\infty} j^2 t_j - \tilde{\mu}_s - 1$$

with t_0 replaced for $\tilde{t_0}S_0$. The latter was however shown not to be very efficient (cf. Sichel 1982b) and a more efficient and equally simple estimate is obtained by matching the mean and the proportion of empty cells to those of the underlying distribution, yielding

$$\tilde{\tau}_{s,2} = 2\tilde{\mu}_s\left(\tilde{\mu}_s + \log\left(\frac{t_0}{C - S_0}\right)\right)\left(\log\left(\frac{t_0}{C - S_0}\right)\right)^{-2} \tag{16}$$

In practice it would however more often be the case that S_0 is unknown and that θ needs to be considered in the estimation process. By employing a zero-truncated approach where only the non-empty sample cell frequencies are considered, this problem is circumvented and θ is accordingly treated as a nuisance parameter since

$$\Pr(f_i = j \mid f_i \geq 1) = \frac{(1-\theta)p_j}{1-(\theta+(1-\theta)p_0)} = \frac{p_j}{1-p_0}, \quad j = 1,2,\ldots . \tag{17}$$

Zero-truncated estimation was described by Sichel (1975, 1982b) who obtained an efficient estimator by matching the average cell size and the proportion of uniques, both amongst the non-empty cells, to the underlying distribution. The estimation procedure entails solving the equation

$$(1 + g)\ln g - Ag + B = 0 \tag{18}$$

for g and where

$$A = \frac{2n}{(C - t_0)} - \ln\frac{n}{t_1} \qquad \text{and} \qquad B = \frac{2t_1}{(C - t_0)} + \ln\frac{n}{t_1}.$$

Equation (18) is easily solved by numerical iteration, e.g. Newton-Raphson, and from the solution $\tilde{g}$ we obtain estimates of μ and τ as

$$\tilde{\mu}_{s,ztr} = \frac{1+\tilde{g}}{2\tilde{g}}\ln\left(\frac{\tilde{g}n}{t_1}\right) \tag{19}$$

and

$$\tilde{\tau}_{s,ztr} = \frac{1-\tilde{g}^2}{2\tilde{g}^2} \tag{20}$$

respectively. An initial estimate to start the iteration is given by the estimates of τ in (16) and then using (20).

5.2. Maximum Likelihood

Maximum likelihood (ML) estimation is fairly straightfoward for the PiG model. When the number of structural zeroes is known a priori, the likelihood is derived from (10) over the C $-S_0$ non-structural-zero cells. Willmot (1987) gave the ML-estimates for the present parameterization and we include them here for the sake of self-containment. The loglikelihood is

$$l_{ML} = \sum_{j=0}^{\infty} t_j \log p_j$$

with t_0 replaced for t_0–S_0 and it is easily shown that the ML-estimate of μ_s is simply the average cell size, i.e.

$$\hat{\mu}_s = \frac{n}{C - S_0} . \tag{21}$$

The ML-estimate of τ_s is the solution to

$$h = \sum_{j=0}^{\infty} t_j \varphi_j - n = 0 \tag{22}$$

with t_0 replaced for t_0–S_0 and where

$$\varphi_j = \frac{(j+1)\, p_{j+1}}{p_j} .$$

The values of φ_j are conveniently computed from the following recursions which are a direct consequence of (6) and (7):

$$\varphi_0 = \frac{\mu}{\eta} \quad \text{and} \quad \varphi_j = \left(\frac{\tau(2j-1)}{\mu^2} + \frac{1}{\varphi_{j-1}} \right) \varphi_0^2, \quad j = 1,2,\ldots .$$

Equation (22) is easily solved using e.g. Newton-Raphson iteration and the required derivative of h with respect to τ_s is

$$\frac{\partial h}{\partial \tau} = \frac{1+\tau_s}{\tau_s} \sum_{j=0}^{\infty} t_j \varphi_j (\varphi_{j+1} - \varphi_j) - \frac{1}{\tau_s} \sum_{j=0}^{\infty} t_j \varphi_j$$

with μ_s replaced by (21). An initial estimate to start the iteration is given by the estimator in (16).

As mentioned in the preceding subsection it would more often be the case that θ is unknown and needs to be considered in the estimation process. By employing a zero-truncated likelihood where only the non-empty sample cell frequencies are considered, θ is treated as a nuisance parameter as shown in (17). This is the approach considered by Skinner and Holmes (1993) designating it a conditional likelihood (CL). The loglikelihood of the f_i for those i which $f_i \geq 1$, is thus defined as

$$l_{CL} = \sum_{j=1}^{\infty} t_j \log \frac{p_j}{1 - p_0} = \sum_{j=1}^{\infty} t_j \log p_j - (C - t_0)) \log(1 - p_0) \tag{23}$$

which yields the system of equations

$$\begin{cases} h_1 = \dfrac{\mu_s}{1-p_0} - \dfrac{n}{C-t_0} = 0 \\ h_2 = \displaystyle\sum_{j=1}^{\infty} t_j \varphi_j - n + \frac{np_0}{\eta_s} = 0 \end{cases} \tag{24}$$

Estimates of μ_s and τ_s are the solutions to (24) and may be obtained by numerical iteration methods such as Newton-Raphson. The derivation of (24) and the required derivatives are provided in the appendix. Small scale experiments indicate however that the rate of convergence for the zero-truncated approach may be slow and that some improvement of the numerical method used may be called for. In our experience using (19) and (20) as initial estimates will usually be a good choice.

5.3. Right-truncation

Skinner and Holmes (1993) also considered truncating the set of probabilities in (18) above a threshold value *m*. The idea is that in applications to disclosure control, a lack of fit in the right hand tail is not likely to be as critical as the left hand tail which may be considered more crucial since only cells belonging to t_0 and t_1 can by definition contain population uniques. A further motivation for this approach is a possible reduction of computational effort. Thus, the p_j are assumed to be proportional to (10) for $j = 1,\ldots,m$ and no assumptions are made about the p_j for $j \geq m+1$. Define

$$q_m^* = \Pr(f_i > m \mid f_i \geq 1) = \frac{1 - \sum_{j=0}^{m} p_j}{1-p_0}$$

The right-truncated version of (23) is then expressed as

$$l_{trCL} = \sum_{j=1}^{m} t_j \log \frac{p_j/(1-p_0)}{1-q_m^*} = \sum_{j=1}^{m} t_j \log p_j - t_m^* \log p_m^*$$

yielding the system of equations

$$\begin{cases} h_1^* = \displaystyle\sum_{j=1}^{m} \frac{jp_j}{p_m^*} - \frac{n_m^*}{t_m^*} = 0 \\ h_2^* = \displaystyle\sum_{j=1}^{m} t_j \varphi_j - n_m^* - \frac{t_m^*}{p_m^*}(m+1)p_{m+1} + \frac{t_m^*}{p_m^*} p_1 = 0 \end{cases} \tag{25}$$

where

$$t_m^* = \sum_{j=1}^{m} t_j, \qquad n_m^* = \sum_{j=1}^{m} j t_j, \qquad \text{and} \qquad p_m^* = \sum_{j=1}^{m} p_j \, .$$

Finding the solutions to (25) requires numerical iteration. E.g. Newton-Raphson requires the derivatives of (25) and it is straightfoward to derive these using the results in the appendix but the result is however not very elegant and convergence may be slow. A further problem indicated by small scale experiments on simulated data seems to be that the truncated approach is sensitive to the choice of starting values in combination with the selected threshold value; depending on the choice the iteration may or may not converge.

As an alternative one might consider the method proposed by Chen and Keller-McNulty (1998) who fit their model to the observed values of t_1 and t_2. Estimators based on their idea, which we will denote by PF12, are thus defined as the solutions to the system of equations

$$\begin{cases} \dfrac{p_1}{1-p_0} = \dfrac{t_1}{C-t_0} \\ \dfrac{p_2}{1-p_0} = \dfrac{t_2}{C-t_0} \end{cases} \tag{26}$$

and is motivated by the same line of reasoning motivating the right-truncation approach. Finding the solutions requires numerical iteration methods such as Newton-Raphson iteration and the required derivatives are straightfoward to derive using the results in the appendix.

5.4. Estimation of θ

The zero-truncated approaches imply estimation of θ. From (9) we have $E(t_0) = C\theta + C(1-\theta)p_0$ so once estimates of μ_s and τ_s are obtained it is straightfoward to estimate θ by replacing $E(t_0)$ for t_0 and p_0 for its estimate, i.e.

$$\hat{\theta} = \frac{t_0 - C\hat{p}_0}{C(1-\hat{p}_0)} \, .$$

As a consequence we have $\hat{t}_0 = t_0$, that is, we obtain a perfect fit for the number of empty cells in the sample. Furthermore, the restriction in (3) is automatically satisfied. For example using (24), the zero-truncated likelihood method yields $\hat{\mu}_s = n(1-\hat{p}_0)/(C t_0)$, and remembering that $\hat{\mu} = N\hat{\mu}_s / n$, it is seen that

$$C(1-\hat{\theta})\hat{\mu} = C\left(1 - \frac{t_0 - C\hat{p}_0}{C(1-\hat{p}_0)}\right)\frac{N(1-\hat{p}_0)}{C-t_0} = N \, .$$

In applications one should be aware of the possibility of obtaining negative estimates of θ which occurs if $t_0 < C\,\hat{p}_0$. This implies that the number of structural

zeroes is negative and indicates that the estimation procedure is over-adjusting to the data. In such cases or if $\hat{\theta}$ is close to zero one may assume that $\theta = 0$ and employ ordinary ML-estimation as described above.

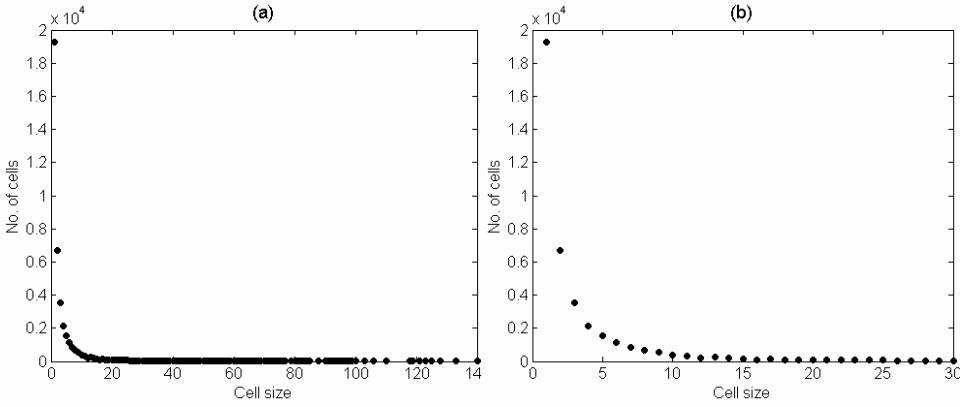
6. Example

6.1. Description of the Data

The data was provided by Statistics Sweden and originates from the 1990 census in Sweden. It consisted of frequency distributions, T_0, T_1, T_2,… for N = 160,536 individuals of ages 20-65 residing in Uppsala county. The following key variables were used: municipal (6), sex (2), age in one-year bands (46), marital status (10), citizenship, Swedish or foreign, (2) and income in 10,000 SEK bands (176). The numbers in parenthesis indicate the number of observed categories of the respective variables from which the total number of combinations is given as C = 1,943,040. Of these T_0 = 1,903,218 were found to be empty leaving a total of 39,822 observed combinations. The number of population uniques T_1 was 19,273 and the largest cell contained 140 units (one cell). Figure 1 displays the population cell size distribution save T_0 and we note the inverse J-shape and the heavy right tail. To illustrate the inverse J-shape in more detail the first 30 cell sizes are also shown.

From the data set a simple random sample without replacement of size n = 16,054 was drawn ($\pi_s = 0.1$). The largest cell size of the 10,046 non-empty cells in the sample was 18 (one cell). The observed number of sample uniques was t_1 = 7,216 or approximately 45% of the sample. Of these 1,952 where found to be true population uniques which is the expected number given the sampling fraction, i.e. approximately 10% of T_1. Thus the ratio (11) defining the risk measure is 0.2705 which is the quantity to be estimated. The sample cell frequencies are provided in table 2.

Figure 1. Population cell size distributions of the example data set. Display (a) shows the entire distribution and (b) shows the details of the left-hand-side. The frequency of empty cells, i.e. T_0, is omitted in both cases.



6.2. Results

Four variations of the PiG-model and methods of estimation were fitted to the data. Given the large number of empty cells it may seem natural to consider only models that truncate at zero but for sake of illustration both the ordinary and the zero-truncated PiG-models were fitted. In both cases we used ML estimation using Newton-Raphson iteration and the moment based estimators (15-16) and (19-20) respectively as starting values. The PF12 estimation procedure for the zero-truncated case was included in the study as well. The equations in (26) were solved by Newton raphson iteration using (19-20) as starting values. We also experimented with the right-truncation method as described above trying various threshold values. Initially we used Newton-Rapson iteration but it was found that better and faster results for this data set were obtained by using the Matlab (2001) minimization routine fminsearch which builds on a simplex direct search method. The routine may however result in local minima so some experimentation with starting values may be required. Here we used the estimators (19-20) without encountering any problems. Choosing the threshold $m = 5$ was found to yield best results in terms of goodness-of-fit measures and when comparing the estimated risk ratio to the true risk ratio.

For comparison two alternative models were considered for the data. The first was Fisher's logarithmic series distribution (LSD). Taking the mixing distribution $g(\lambda)$ in (2) to be a gamma distribution results in the negative-binomial (NB) distribution. In many cases it has however been noted that the α parameter of the NB tends to be very small in disclosure applications. In such cases it may be appropriate to consider instead the limiting distribution of the zero-truncated NB as $\alpha \to 0$ which results to the LSD. The pdf of the LSD is defined by

$$\Pr(F_i = j) = -\frac{\phi^j}{j \log(1-\phi)}, \quad j = 1, 2, \ldots$$

where $0 < \phi < 1$. Assuming Bernoulli sampling the marginal distribution of the cell sizes is also LSD with parameter

$$\phi_s = \frac{\pi_s \phi}{1 - \phi(1 - \pi_s)}$$

It can be shown that R_2 is simplified to

$$R_2 = -\frac{n}{C - t_0} \frac{(1-\phi)\log(1-\phi_s)}{\phi_s}.$$

The LSD model was fitted using ordinary ML estimation and Newton-Raphson iteration.

The second alternative model was the zero-truncated Poisson-lognormal (PLN) distribution proposed by Skinner and Holmes (1993, 1998). The model is defined by choosing $g(\lambda)$ in (2) to be distributed as lognormal with parameters μ and $\sigma^2 < 0$. Assuming Bernoulli sampling the marginal distribution of the cell sizes is also PLN with parameters $\mu_s = \mu + \log\pi_s$ and $\sigma_s^2 = \sigma^2$. Unfortunately the PLN distribution is not available in closed form so numeric integration is required to calculate the probabilities and the risk measure R_2. For this data set we experimented with various variable substitutions of the lognormal kernel and different numeric integration techniques and settled for the transformation $\lambda = (1–t) / t$ to obtain finite integration limits and the Matlab (2001) quadl routine which uses an adaptive quadrature technique. Skinner and Holmes suggested either censoring or truncating the loglikelihood above a threshold value m. We tried both methods on the sample data and found that choosing $m = 4$ for the censored version and $m = 5$ for the right-truncated version yielded best results in terms of goodness-of-fit measures and in comparison to the true risk ratio. Maximizing the censored and truncated loglikelihoods also required some experimentation including Newton-Raphson and the Nelder-Mead method mentioned above. Both methods were found to be sensitive to the choice of starting values and the latter occasionally produced negative estimates of σ^2.

To compare the fit to the different models two conventional goodness-of-fit statistics, the Pearson statistic

$$\chi_P^2 = \sum \frac{(t_j - \hat{t}_j)^2}{\hat{t}_j}$$

and the likelihood-ratio statistic (LRT)

$$\chi_{LR}^2 = 2\sum t_j \log(t_j / \hat{t}_j)$$

were calculated for each model. Both statistics were modified in the obvious way when categories were collapsed. The results are summarized in tables 1 and 2.

Table 1. Estimates of model parameters, number of population uniques T_1, and risk measure R_2 and loglikelihood.

Model	Parameters			Loglikel	$\hat{T}_1$	$\hat{R}_2$
Logarith. series dist., ML	$\hat{\phi}_s = 0.583$			-5169.0	10724	0.1601
Poisson-lognormal	$\hat{\mu}_s =$	$\hat{\sigma}_s^2 =$	$\hat{\theta} =$			
(1) z-tr, cens. $m = 4$, ML	-3.331	3.247	0.951	-9253.7	16646	0.2306
(2) z-tr, r-tr, $m = 5$, ML	-3.622	3.657	0.945	-8206.2	17366	0.2419
Poisson-inverse Gaussian	$\hat{\mu}_s =$	$\hat{\tau}_s =$	$\hat{\theta} =$			
(1) ML	0.008	1.893	-	-72972.4	25286	0.3448
(2) z-tr, ML	0.074	1.750	0.889	-10058.7	21636	0.2999
(3) z-tr, PF12	0.117	1.552	0.931	-10062.4	19629	0.2720
(4) z-tr, r-tr, $m = 5$, ML	0.106	1.476	0.924	-8207.9	20348	0.2793

Our first remark is that the LSD performs badly with this data set, both in terms of fitting to the data as measured by the χ^2 statistics and in predicting the risk ratio and the total number of population uniques. This is not surprising as it agrees with the results of previous studies, e.g. Skinner et al. (1994), Chen and Keller-McNulty (1998) and Hoshino (2001). The fitted values of t_1 and t_2 show a poor fit to the observed values and the decay of the right hand tail appears to rapid. The resulting estimates of R_2 and T_1 are accordingly not satisfactory.

The PLN and the PiG models, save PiG (1), on the other hand both appear to adapt better to the frequency structure. The poor results of the PiG (1) is apparently a result of ignoring the large number of empty cells and assuming that $\theta = 0$. It appears as if most of the effort in fitting to the data is waisted on strectching out to t_0 on the expense of the other cell size frequencies. Even so, compared to the LSD even the PiG (1) model performs surprisingly well. When the zero-truncated methods are used better results are obtained both in fit to the data and in predicting R_2 and T_1. It is interesting to note that the best results with respect to predicting R_2 and T_1 are obtained when the estimation procedure is focused on the small cell sizes as in PiG (3) which is the PF12 estimation method. This seems to corroborate with the results of Chen and Keller-McNulty (1998). Furthermore, as mentioned in the preceding subsection, the censoring and trunction thresholds of the PLN (1) and (2) and the PiG (4) were opportunisticly chosen to produce estimates close to the true value of R_2. We found that higher thresholds for the PLN increasingly underestimated R_2 while for the PiG (4), R_2 was increasingly overestimated.

Table 2. Observed and fitted cell size frequencies and goodness-of-fit statistics for sample data set. The (*) indicates collapsing of categories above and including the corresponding cell size. The models are LSD: logarithmic series distribution and ML estimation, PLN: (1) zero-truncated and censored likelihood, $m = 4$, (2) zero- and right-truncated likelihood, $m = 5$. PiG: (1), full likelihood,

θ = 0, (2), zero-truncated likelihood, (3), zero-truncated and the PF12 estimator, (4), zero- and right-truncated likelihood, $m = 5$.

Size	Observ.	Fitted $\hat{t}_j$						
		LSD	PLN		PiG			
j	t_j		(1)	(2)	(1)	(2)	(3)	(4)
0	1932994	-	-	-	1932993.2	-	-	-
1	7216	6697.2	7217.7	7220.3	7300.8	7216.5	7216	7218.3
2	1573	1951.7	1561.4	1550.4	1457.6	1529.5	1573	1540.0
3	533	758.3	555.7	562.7	576.5	596.3	598.8	578.6
4	272	331.5	258.0	267.2	285.0	290.0	283.5	270.5
5	155	154.6	453.2*	148.4	157.8	157.9	150.2	141.5
6	117	75.1	-	-	93.6	92.1	85.2	-
7	70	37.5	-	-	58.2	56.3	50.6	-
8	41	19.1	-	-	37.4	35.6	31.1	-
9	36	9.9	-	-	24.7	23.1	19.6	-
10	11	5.2	-	-	16.6	15.2	12.6	-
11	8	2.8	-	-	11.3	10.2	8.2	-
12	4	1.5	-	-	7.8	7.0	5.5	-
13	5	1.7*	-	-	5.5	4.8	3.6	-
14	3	-	-	-	3.9	3.3	2.5	-
15	1	-	-	-	2.8	2.3	1.7	-
16	0	-	-	-	2.0*	5.8*	3.8*	-
17	0	-	-	-	-	-	-	-
18	1	-	-	-	-	-	-	-
19+	0	-	-	-	-	-	-	-
Pearson χ^2		396.74	1.78	2.28	39.39	34.96	47.46	5.60
LRT χ^2		338.84	1.78	2.30	42.38	36.07	43.58	5.65
d.f.		11	2	2	14	13	13	2

7. Remarks

Since the scope of this paper has been limited to a theoretical review with only a small-scale example, it is necessarily difficult to evaluate how the PiG model fares in general and when compared to alternative approaches. Before any conclusions can be made a more extensive evaluation is called for including tests on real-life data and comparisons with other models. Such an evaluation is intended to appear in a separate report. The PiG model does however provide an analytically tractable alternative and calculations of the disclosure risk along the lines discussed are easily computed.

In the present paper we have only considered the two-parameter version of the more general three-parameter PiG, commonly known as the Sichel distribution. It is defined by

$$P_j = \frac{\eta^{-\gamma}}{j!}\left(\frac{\mu}{\eta}\right)^j \frac{K_{\gamma+j}(\mu\eta/\tau)}{K_\gamma(\mu/\tau)}$$

where $-\infty < \gamma < \infty$. This distribution was introduced by Sichel (1971) and the distribution in (5) is obtained by setting $\gamma = -1/2$. A short review of the Sichel distribution is given in Johnson et al. (1992, pp. 455-457). This three-parameter distribution is very powerful and a number of known distribution functions such as the Poisson, negative binomial, geometric, Fisher's logarithmic series, are special or limiting forms of the Sichel. A problem with the Sichel distribution is however that the derivative $(\partial/\partial\gamma)K_\gamma(z)$ is not available in closed form and the estimation of γ requires special attention, cf. Stein et al (1987).

We note also that the risk measures in (14) provides only an overall measure of disclosure risk pertaining to the sample as a whole. A per-record measure of disclosure risk is perhaps more useful as it would provide a means to identify sensitive (unique) records to which disclosure controlling measures can be applied. From an intruders point of view it would be optimal to utilize as much as possible of the information provided in the sample when formulating a model. Methods which attempt to capture the underlying probability structure inherent from the key variables defining X have been suggested, see e.g. Fienberg and Makov (1998) and Skinner and Holmes (1998). The latter considered a per-record measure based on their Poisson-lognormal model and we note that similar regression methods are available for PiG data based on a model of the form $\mu_{\mathbf{x}} = \exp(\mathbf{x}'\beta)$ with τ fixed, see Dean et al. (1989) for details. Furthermore, as pointed out by an anonymous referee, the problem can also be addressed from a Bayesian viewpoint. In the Nordic European countries detailed population statistics are frequently being published from registers and population uniques can either be inferred or excluded directly from the published tables or the published tables can be used as auxiliary information along with the sample data. In conclusion, the possibility of extending the present model in these directions is certainly worthy of future exploration.

Acknowledgement

The author wishes to thank Professor Daniel Thorburn, Dep. of Statistics, Stockholm University, for his many comments and suggestions in the preparation of this paper and Hans Block, Statistics Sweden, for providing the data set used in the example. The author also acknowledges the valuable comments of an anonymous referee.

Appendix. Derivation of Likelihood Equations

In the following the index *s* on the parameters, indicating sample level, is dropped for notational ease. The first derivatives of the log probabilities with respect to the parameters are (see Willmot, 1987, for details)

$$\frac{\partial \log p_j}{\partial \mu} = \frac{1}{\tau} + \frac{2j}{\mu} - \frac{\eta^2}{\mu\tau}\varphi_j \tag{27}$$

and

$$\frac{\partial \log p_j}{\partial \tau} = -\frac{\mu}{\tau}\left(\frac{\partial \log p_j}{\partial \mu}\right) + \frac{j}{\tau} - \frac{\varphi_j}{\tau} \tag{28}$$

where $\varphi_j = (j+1)p_{j+1}p_j^{-1}$ from which it in turn is easy to derive that

$$\frac{\partial p_j}{\partial \mu} = \frac{1}{\tau}p_j + \frac{2}{\mu}jp_j - \frac{\eta^2}{\mu\tau}(j+1)p_{j+1}$$

and

$$\frac{\partial p_j}{\partial \tau} = -\frac{\mu}{\tau}\left(\frac{\partial p_j}{\partial \mu}\right) + \frac{1}{\tau}jp_j - \frac{1}{\tau}(j+1)p_{j+1}.$$

Furthermore we need

$$\frac{\partial}{\partial \mu}\log(1-p_0) = -\frac{1-\eta}{\tau}\frac{p_0}{1-p_0}$$

and

$$\frac{\partial}{\partial \tau}\log(1-p_0) = -\frac{\mu}{\tau}\frac{(1+\tau-\eta)}{\tau\eta}\frac{p_0}{(1-p_0)}.$$

For the second derivatives we will need also the derivatives of φ_j and it is straightfoward using (27) and (28) to show that

$$\frac{\partial \varphi_j}{\partial \mu} = \frac{2}{\mu}\varphi_j - \frac{\eta^2}{\mu\tau}\varphi_j(\varphi_{j+1} - \varphi_j)$$

and

$$\frac{\partial \varphi_j}{\partial \tau} = -\frac{1}{\tau}\varphi_j + \frac{1+\tau}{\tau^2}\varphi_j(\varphi_{j+1} - \varphi_j).$$

In the following derivation of the likelihood equations we use the same line of arguments as Willmot (1987) and as an example we consider the conditional log

likelihood in (23); the other cases are analogous. The first derivatives of (23) with respect to μ and τ are

$$\frac{\partial l_{CL}}{\partial \mu} = \sum_{j=1}^{\infty} t_j \frac{\partial \log p_j}{\partial \mu} - (C - t_0) \frac{\partial \log(1 - p_0)}{\partial \mu} \tag{29a}$$

$$= \frac{C - t_0}{\tau} + \frac{2n}{\mu} - \frac{\eta^2}{\mu\tau} \sum_{j=1}^{\infty} t_j \varphi_j + (C - t_0) \frac{(1-\eta) p_0}{\tau(1 - p_0)} \tag{29b}$$

and

$$\frac{\partial l_{CL}}{\partial \tau} = \sum_{j=1}^{\infty} t_j \frac{\partial \log p_j}{\partial \mu} - (C - t_0) \frac{\partial \log(1 - p_0)}{\partial \mu} \tag{30a}$$

$$= -\frac{\mu}{\tau} \sum_{j=1}^{\infty} t_j \frac{\partial \log p_j}{\partial \mu} + \frac{\eta}{\tau} - \frac{1}{\tau} \sum_{j=1}^{\infty} t_j \varphi_j + (C - t_0) \frac{\mu(1 + \tau - \eta) p_0}{\tau^2 \eta (1 - p_0)} \tag{30b}$$

respectively. It is clear that the partials of (23) with respect to μ and τ are identically zero when the likelihood is maximized, i.e. at the CL-estimates $\hat{\mu}$ and $\hat{\tau}$. Thus from (29a) we have that

$$\sum_{j=1}^{\infty} t_j \frac{\partial \log p_j}{\partial \mu} \bigg|_{\mu=\hat{\mu}, \tau=\hat{\tau}} = (C - t_0) \frac{\partial \log(1 - p_0)}{\partial \mu} \bigg|_{\mu=\hat{\mu}, \tau=\hat{\tau}} \tag{31}$$

and it follows from setting (30b) equal to zero and using (31) that

$$\sum_{j=1}^{\infty} t_j \hat{\varphi}_j = n - \frac{\hat{\mu}(C - t_0)}{\hat{\eta}} \frac{\hat{p}_0}{1 - \hat{p}_0} \tag{32}$$

where $\hat{\varphi}_j$, $\hat{\eta}$ and $\hat{p}_0$ are the CL-estimates of φ_j, η and p_0 respectively. Thus, setting (29) and (30) equal to zero and using (32) in (29b) yields after simplification the first likelihood equation h_1 in (24). The second equation h_2 is simply (32) with μ replaced by $n(1-p_0)(C-t_0)^{-1}$ from the first equation. It is straightfoward using the results above to show that the required derivatives of h_1 and h_2 are

$$\frac{\partial h_1}{\partial \mu} = \frac{1}{1-p_0} + \frac{\mu(1-\eta)}{\tau}\frac{p_0}{(1-p_0)^2}$$

$$\frac{\partial h_1}{\partial \tau} = \frac{\mu^2(1+\tau-\eta)}{\tau^2\eta}\frac{p_0}{(1-p_0)^2}$$

$$\frac{\partial h_2}{\partial \mu} = \frac{2}{\mu}\sum_{j=1}^{\infty} t_j\varphi_j - \frac{\eta^2}{\mu\tau}\sum_{j=1}^{\infty} t_j\varphi_j(\varphi_{j+1}-\varphi_j) - \frac{n(1-\eta)p_0}{\tau\eta}$$

$$\frac{\partial h_2}{\partial \tau} = -\frac{1}{\tau}\sum_{j=1}^{\infty} t_j\varphi_j + \frac{1+\tau}{\tau^2}\sum_{j=1}^{\infty} t_j\varphi_j(\varphi_{j+1}-\varphi_j) - \frac{np_0}{\eta}\left(\frac{\mu(1+\tau-\eta)}{\tau^2\eta} - \frac{1}{\eta^2}\right).$$

REFERENCES

ABRAMOVITZ, M. and STEGUN, I.A. (1970) *Handbook of Mathematical Functions*. Monograph. New York: Dover Publications.

BETHLEHEM, J.G., KELLER, W.J. and PANNEKOEK, J. (1990) Disclosure Control of Microdata. *Journal of the American Statistical Association*, 85, pp. 38-45.

BLIEN, U., MÜLLER, W. and WIRTH, H. (1993) Needles in Haystacks Are Hard to Find – Testing Disclosure Risks of Anonymous Individual Data. In *Proceedings of the International Seminar on Statistical Confidentiality, Dublin, 8-10 September 1992*, pp. 391-406. Luxembourg: Off. for Official Publications of the European Commun.

BLOCK, H and OLSSON, L. (1976) Backwards Identification of Personal Information – Bakvägsidentifiering (in Swedish). *Statistisk Tidskrift*, 4, pp. 135-144.

CHEN, G. and KELLER-MCNULTY, S. (1998) Estimation of Identification Disclosure Risk in Microdata. *Journal of Official Statistics*, 14, pp. 79-95.

DALENIUS, T. (1977) Towards a Methodology For Statistical Disclosure Control. *Statistisk Tidskrift*, 5, pp. 429-444.

DALENIUS, T. (1986) Finding a Needle In a Haystack or Identifying Anonymous Census Records. *Journal of Official Statistics*, 2, pp. 329-336.

DEAN, C., LAWLESS, J.F. and WILLMOT, G.E. (1989) A mixed Poisson-inverse-Gaussian regreesion Model. *The Canadian Journal of Statistics*, 17, pp. 171-181.

DOMINGO-FERRER, J. (ed.) (2002) *Inference Control in Statistical Databases*. Monograph. Berlin: Springer.

DOYLE, P., LANE, J., THEEUWES, J.J.M., and ZAYATZ, L. (eds.), (2001) *Confidentiality, Disclosure, and Data Access*. Monograph. Amsterdam: Elsevier.

DUNCAN, G. and LAMBERT, D. (1989) The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics*, 7, pp. 207-217.

DUNCAN, G.T. and PEARSON, R.W. (1991) Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future. With discussion. *Statistical Science*, 6, pp. 219-239.

ELLIOT, M.J., SKINNER, C.J. and DALE, A. (1998) Special Uniques, Random Uniques and Sticky populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. *Research in Official Statistics*, 1, pp. 53-67.

FIENBERG, S.E. and MAKOV, U. (1998) Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data, *Journal of Official Statistics*, 14, pp. 385-397.

FOLKS, J.L. and CHHIKARA, R.S. (1978) The Inverse Gaussian Distribution and Its Statistical Application – A Review. With discussion. *Journal of the Royal Statistical Society, Series B*, 40, pp. 263-289.

FRANK, O. (1976) Individual Disclosures from Frequency Tables. In T. Dalenius and A. Klevmarken (eds.) *Personal Integrity and the Need for Data in the Social Sciences*. Swedish Council for Social Science Research, pp. 175-187.

FRANK, O. (1988) Designing Classifiers for Partial information Release, in H.H. Bock (editor) *Classification and related methods of data analysis: proceedings of the First Conference of the IFCS, Tech. Univ. of Aachen,* pp. 687-690. New York: North-Holland.

GREENBERG, B.V., ZAYATZ, L.V. (1992) Strategies for Measuring Risk in Public Use Microdata Files. *Statistica Neerlandica*, 46, pp. 33-48.

HOLLA, M.S. (1966) On a Poisson-inverse Gaussian Distribution, *Metrika*, 11, pp. 115-121.

HOSHINO, N. (2001) Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, 17, pp. 499-520.

HOSHINO, N. and TAKEMURA, A. (1998) On the Relation Between Logarithmic Series Model and Other Superpopulation Models Useful for Microdata Disclosure Risk Assessment. Discussion Paper, 98-F-7, Faculty of Economics, University of Tokyo. (Published in *Journal of Japan Statistical Society*, 28, pp. 125-134, 1998, after revision).

JOHNSON, N.L., KOTZ, S. and BALAKRISHNAN, N. (1994) *Continuous Univariate Distributions*, Vol. 1, 2nd ed.. Monograph. John Wiley & Sons, New York.

JOHNSON, N.L., KOTZ, S. and KEMP, A.W. (1992) *Univariate Discrete Distributions*, 2nd ed.. Monograph. John Wiley & Sons, New York.

LAMBERT, D. (1993) Measures of Disclosure Risk and Harm. *Journal of Official Statistics*, 9, pp. 313-331.

Matlab (2001) Matlab Ver. 6.1, Release 12.1, MathWorks Inc..

ORD, J.K. and WHITMORE, G. (1986) The Poisson-Inverse Gaussian distribution as a model for species abundance, *Communications in Statistics - Theory and Methods*, 15, pp. 853-871.

PAASS, G. (1988) Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business & Economic Statistics*}, 6, pp. 487-500.

SAMUELS, S.M. (1998) A Bayesian, Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, 14, pp. 373-383.

SÄRNDAL, C.E., SWENSSON, B. and WRETMAN, J. (1992) *Model Assisted Survey Sampling*. Monograph. New York: Springer-Verlag.

SICHEL, H.S. (1971) On a Family of Discrete Distributions Particularly Suited to Represent Long-tailed Frequency Data. In N.F. Laubscher (editor*) Proceedings of the Third Symposium on Mathematical Statistics*, pp. 51-97. Pretoria: C.S.I.R.

SICHEL, H.S. (1973) The Density and Size Distribution of Diamonds*, Bulletin of the International Statistical Institute*, 45 (2), pp. 420-427.

SICHEL, H.S. (1974) On a Distribution Representing Sentence-length in Written Prose, *Journal of the Royal Statistical Society, Series A*, 137, pp. 25-34.

SICHEL, H.S. (1975) On a Distribution Law for Word Frequencies, *Journal of the American Statistical Association*, 70, pp. 542-547.

SICHEL, H.S. (1982a) Repeat-buying and the Generalized Inverse Gaussian-Poisson Distribution, *Applied Statistics*, 31, pp.193-204.

SICHEL, H.S. (1982b) Asymptotic Efficiencies of Three Methods of Estimation for the Inverse Gaussian-Poisson Distribution, *Biometrika*, 69, pp. 467-472.

SKINNER, C.J. and HOLMES, D.J. (1993) Modelling Population Uniqueness. In *Proceedings of the International Seminar on Statistical Confidentiality, Dublin, 8-10 September 1992*, pp. 175-199. Luxembourg: Office for Official Publications of the European Communities.

SKINNER, C.J. and HOLMES, D.J. (1998) Estimating the Re-identification Risk Per Record. *Journal of Official Statistics*, 14, pp. 361-372.

SKINNER, C., MARSH, C., OPENSHAW, S. and WYMER, C. (1994) Disclosure Control for Census Microdata, *Journal of Official Statistics*, 10, pp. 31-51.

ST-CYR, P. (1998) Modelling Population Uniqueness Using a Mixture of Two Distributions. In *Statistical Data Protection – Proceedings of the Conference, Lisbon, 25 to 27 March 1998-1999 edition*, pp. 277-286. Luxembourg: Office for Official Publications of the European Communities.

STEIN, G., ZUCCHINI, W. and JURITZ, J.M. (1987) Parameter Estimation for the Sichel Distribution and its Multivariate Extension, *Journal of the American Statistical Association*, 82, pp. 938-944.

TAKEMURA, A. (1999) Some Superpopulation Models for Estimating the Number of Population Uniques. In *Statistical Data Protection - Proceedings of the Conference, Lisbon, 25 to 27 March 1998-1999 edition*, pp. 59-76. Luxembourg: Office for Official Publications of the European Communities.

WILLENBORG, L. and de WAAL, T. (1996*) Statistical Disclosure Control in Practice; Series: Lecture Notes in Statistics*, Vol. 111. Monograph. Springer-Verlag, New York.

WILLENBORG, L.C. and de WAAL, T. (2000*) Elements of Statistical Disclosure Control; Series: Lecture Notes in Statistics*, Vol. 155. Monograph. New York: Springer-Verlag.

WILLMOT, G.E. (1987) The Poisson-inverse Gaussian Distribution as an Alternative to the Negative Binomial, *Scandinavian Actuarial Journal*, pp. 113-127.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 927–942

THE DANISH HEALTH AND MORBIDITY SURVEY 2000 – DESIGN AND ANALYSIS

Michael Davidsen[1], Mette Kjøller[2]

ABSTRACT

Introduction: In several countries health interview surveys are conducted regularly to monitor health status of populations. A health and morbidity survey programme was initiated in Denmark in1987 to monitor health status and describe changes over time. In the year 2000 survey, a further crucial aim was to serve as a tool for Danish counties in their health care planning. This paper describes how the objectives were combined into one survey, the weighting scheme used, the analysis and basic reporting of results.

Material and methods: The sample consists of three subsamples. These include a national, a follow-up and a supplementary county sample, ensuring at least 1,000 persons with a completed interview in each county. Weights are constructed taking into account different county distributions and different questions asked in the various subsamples. The basic parameters to be estimated were prevalence (proportions) and odds-ratio. A logistic regression model was used to estimate odds-ratios with 95% confidence intervals in groups defined by socio-demographic variables.

Results: 22,486 persons were selected for the sample and 74.2% completed an interview. The number of interviews in each county exceeded 1,000 by up to 47%. The sample seems representative of the adult Danish population. An example of the presentation of results is given and described in detail. The odds-ratios provide an age- and sex-adjusted comparison between groups. Changes over time are shown in total and in groups defined by age and sex.

Discussion: The way national health interview surveys are conducted in Europe varies across countries. Odds-ratios based on logistic regression are chosen, as they are a natural and statistically attractive measure, which is easy to interpret. The difference between SAS and SUDAAN are expected to be small due to small sampling fractions.

[1] National Institute of Public Health, Svanemøllevej 25, DK-2100 Copenhagen; e-mail: md@niph.dk

[2] National Institute of Public Health, Svanemøllevej 25, DK-2100 Copenhagen.

Key words: Health Interview Survey /logistic regression/ stratified sample.

1. Introduction

In Europe, as well as in countries all over the world, there is a need to monitor population health. One way of describing the health of a population is by means of official statistical registers, e.g. cause of death registers. Another way is by means of health interview surveys, which give the possibility of describing health and morbidity both in a daily life perspective and by factors influencing it such as lifestyle, risk factors and living conditions. Such surveys are conducted on a more or less regular basis all over the world [U.S. Department 1999]. In Europe, several countries perform National Health Interview Surveys [U.S. Department 1999; Hupkens 1999]. While the specific purpose of the surveys varies, the monitoring of health status is the most common reason.

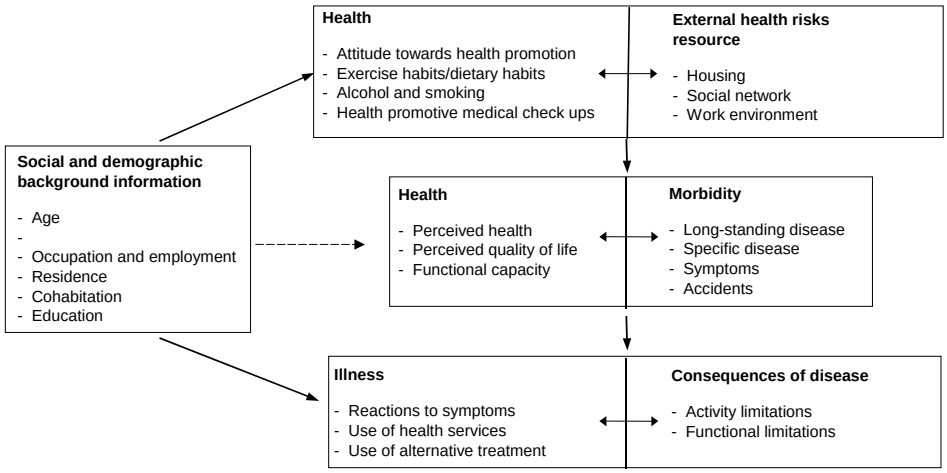


Figure 1. Core elements of the Danish Health Interview Survey Programme

In Denmark, a National Health and Morbidity Survey programme was initiated in 1987. Two of the main objectives were to describe health status, including health behaviour and health habits, lifestyles, social network, working conditions and living conditions, as well as to describe changes in health and morbidity over time. A main purpose is to describe changes over time and, therefore, the programme is structured around a set of core elements unaltered over time. Additional topics are included in the surveys due to either current

discussion on health policy or awareness generated by health professionals. The core elements of the programme are structured in a traditional epidemiological model describing risk factors that might lead to ill health and morbidity, as well as the consequences of ill health (figure 1). The arrows between the boxes suggest causal relationships between the various elements. The model serves as a tool for understanding and interpretation.

As a result of this programme, two cross-sectional health interview surveys have been conducted, one in 1987 [Rasmussen 1988] and one in 1994 [Kjøller 1995]. The target population in both these surveys was adult Danish citizens (age 16 or more) and the sample design was simple random sampling. Data were collected via personal interview at the respondents' home. Following the interview, respondents in the 1994 survey were asked to complete a self-administered questionnaire. Data collection was divided into three rounds, i.e. winter, spring and autumn.

Within the Health and Morbidity Survey programme, a new survey was carried out in 2000 [Kjøller 2002]. In addition to the already stated purposes, a crucial aim with this survey was to provide data for local health care planning. In Denmark, the counties are responsible for hospital services, specialist services and general practitioners services. According to law, each county establishes a health care plan every fourth year. Furthermore, the survey should serve as a panel-study of the 1994 survey primarily in order to obtain longitudinal data to describe development in health and morbidity of the Danish population.

The aim of the present paper is to describe the design and analysis of the health and morbidity survey year 2000. This includes the sampling procedure used to combine the objectives of the survey, the weighting scheme, the statistical analysis and the presentation of results.

2. Material and methods

Denmark is a rather small country divided into 15 counties of varying sizes (between 35,000 and 490,000 persons). In 2000, there were 4.1 million adult (age 16 or more) Danes.. Each Dane has a unique personal registration number and is registered in the Civil Person Register. This register contains information on, among other things, gender, age, address, citizenship and county of residence. It is a dynamic register, as all previous and present information on every person is kept.

2.1. Sample selection

The sample for the 2000 Danish Health and Morbidity Survey consists of three subsamples as follows:

The *national sample* is a close parallel to the previous two surveys both in design and in size. The sample was selected by stratification proportional to county size.

The *follow-up sample* is intended to be both a panel of the survey in 1994 and a sample of adult Danes. Therefore, it consists of three parts: (a) all persons invited to the survey in 1994 and still alive in 2000 (b) a supplement of persons aged 16-21 in year 2000. This sample was chosen stratified proportional to county size. (c) a supplement of persons who have obtained Danish citizenship between 1994 and 2000. In this part of the follow-up sample simple random sampling was used.

The ***supplementary county sample*** ensures that the total sample is large enough to be practical for the Danish counties in their health care planning. Thus, the purpose of this sample was to ensure at least 1,000 completed interviews in each county. As one county (Bornholm) is considerably smaller than the others, it was decided that 600 completed interviews would be sufficient in this county.

As in the previous surveys, the primary sampling unit was individual persons, i.e. Danish citizens aged 16 or more. All persons were randomly selected. Data were collected as in previous surveys, i.e. collection in three rounds. The sampling frame used was based upon the Civil Person Register and conducted by SFI-SURVEY, a division of the Danish National Institute of Social Research. The order of selection of the samples was as follows: all persons from the 1994 survey, the national sample, young persons from the follow-up sample, and the supplementary county sample. Part (c) of the follow-up sample (immigrants) was treated separately because it had to be drawn among persons who had obtained Danish citizenship between 1994 and 2000. The chosen procedure ensured that each person could only be selected to one subsample.

As the aim was to have at least 1,000 completed interviews, the sample size depended upon the response rate in each county. The third and last round of data collection was used to ensure that this goal was achieved.

A questionnaire was developed for each subsample. The core elements were included in all questionnaires while questions on other issues were referred to different combinations of subsamples.

The data collection was as in the previous surveys [Rasmusen 1987, Kjøller 1994] and briefly described in the introduction.

2.2. Weighting

The design chosen implies that the county distribution in the total sample is different from the county distribution in the country as a whole. In developing the weighting system, it was considered essential to take this fact into account. Also taken into account was the fact that different questions were asked in different combinations of subsamples and that both interview and self-administered questionnaires were used.

First, the design weights were defined. These were chosen as in a county stratified sample. That is, we use:

$$w_D = \frac{N_c}{n_c}$$

where N_c is the total population size and n_c is the number of selected persons in the total sample in county c (=1,…,15). Thus, as the sample design is considered stratified by county, all persons in a county receive the same weight.

In order to take into account the type of questionnaire (interview or self-administered) the fact that some questions were asked in some samples only, post-stratification was used in the following way: Let A denote the national sample, F the follow-up sample and U the supplementary county sample and let r be one of the elements in the set

{U, AU, FU, AFU}. Let n_c^r be the number of invited persons in sample r and county c, and let m_c^r be the number of persons who fulfilled an interview. Observe that $n_c = n_c^{AFU}$

We then define

$$w_a^r = \frac{n_c}{n_c^r}$$

and

$$w_b^r = \frac{n_c}{m_c^r}$$

For questions asked in interview in sample r, we then define

$$w_I^r = w_D * w_a^r \quad (1)$$

and for questions asked in self-administered questionnaires in sample r, we define

$$w_S^r = w_D * w_b^r \quad (2)$$

In this way, eight weights were defined. All weights were normalized by dividing with the average weight in the appropriate population. That is, when applying the weight w_I^r it was ensured that the weights summed to the number of persons having an interview in sample r. Likewise, when applying the weight w_S^r, they sum to the number of persons returning a self-administered questionnaire in sample r.

Questions were also asked in the national and all follow-ups, in this instance a normalized weight of 1 was used.

Item non-response in the interview is very small (in most questions, not more than 1%).

No weighting was used in the 1987 and 1994 surveys.

2.3. Statistical method

The analysis here described the aim of presenting results in the final report of the survey [Kjøller 2002]. The first step was to identify approximately 170 indicators mostly from the core elements. Each indicator was calculated on the basis of information from one or more questions posed and dichotomised into 'yes' and 'no', with the presentation of results focusing on 'yes'

For each indicator, the prevalence (proportions), odds-ratios with 95% confidence intervals, and number of persons were calculated in groups defined by socio-demographic variables. The socio-demographic variables used were age (16-24, 25-44, 45-66, 67-79 and 80+ years) and sex (men/women) combined into 10 groups, education according to the International Standardized Classification of Education (ISCED – 5 categories plus a category for school attendants and persons not being classifiable), socio-economic status (13 categories), cohabitation status (5 categories) and county (15 categories).

Odds-ratios (OR) were used to make group comparisons of the socio-demographic variable controlling for age and sex. Odds-ratios with 95% confidence intervals were estimated by logistic regression [Hosmer 2000]. For age and sex we used the model

$$\text{logit}(p) = \text{age}*\text{sex}$$

where p is the probability of a person answering 'yes' to the indicator and age and sex are categorical variables. This model implies that the pattern seen by prevalence and odds-ratio are the same.

For each of the other socio-demographic variables (SDV), we used the model

$$\text{logit}(p) = \text{SDV} + \text{age}*\text{sex}$$

where p is defined as above and, again, treating the socio-demographic variable and age and sex as categorical. These models were chosen in order to compare odds-ratios across groups defined by the socio-demographic variable controlling for age and sex.

With the exception of county, a fixed reference group was chosen for all socio-demographic variables. For county, odds-ratios were constructed as deviates from a country average, i.e. an average of the odds in each county [Hosmer 2000].

Maps indicating crude grouping of counties were constructed if the Wald-test assuming equal county odds showed significance at the 5% level. The map displays county prevalence grouped by dividing the interval from the lowest to the highest prevalence into three equally long (equidistant) intervals, regardless of the number of counties in each group.

In the description and analysis of changes over time, the surveys in 1987, 1994 and 2000 were regarded as independent cross-sectional surveys. The statistical testing of changes over time was based on logistic regression including age, sex and year of survey. Backward elimination starting with the (hierarchical) model

$$\text{logit}(p) = \text{year} * \text{sex} * \text{age}$$

was conducted to see if any effect of year was seen (sex and age were always included in the analysis). The level of significance was 5%.

Number of persons was calculated on the basis of the sample(s) in which the question underlying the dichotomous indicator was asked.

Prevalence was calculated using the weighting described earlier. All logistic regressions were conducted using logistic regression with a WEIGHT-statement in SAS 8.2 [SAS 1999],

When analysing changes over time, a weight of one was used for 1987 and 1994.

3. Results

3.1. Design

Table 1 shows the population size, number of invited and interviewed persons, the sampling fraction (number of invited / population size) and the response rate in each county, sorted according to population size. The goal of at least 1,000 interviews in each county except Bornholm was fulfilled. The number of interviewed persons exceeds 1,000 by up to 47% indicating no firm stopping rule for inclusion of new persons. As seen, the sampling fraction is rather small (0.55%) ranging between 0.37% in the largest county (Aarhus) to 2.4% in the smallest. All in all, 22,486 persons were selected and 16,690 completed an interview, giving a response rate of 74.2%. The response rate is lowest in Copenhagen City and Copenhagen County, i.e. the capital area of Denmark.

Table 1. County specific population size, invited, sampling fraction, interviewed and response rate in the Danish Health and Morbidity Survey 2000

County	Population size (1000)	Number of invited	Sampling fraction (%)	Number of interviewed	Response rate (%)
Aarhus	490	1814	0.37	1474	81.3
Copenhagen County	460	2063	0.45	1380	66.9
Copenhagen City	451	2020	0.45	1293	64.0
Nordjylland	386	1534	0.40	1188	77.4
Fyn	368	1469	0.40	1119	76.2

County	Population size (1000)	Number of invited	Sampling fraction (%)	Number of interviewed	Response rate (%)
Frederiksborg	275	1441	0.52	1100	76.3
Vejle	268	1232	0.46	1032	83.8
Vestsjaelland	229	1450	0.63	1022	70.5
Ringkøbing	209	1374	0.66	1090	79.3
Storstrøm	206	1442	0.70	1014	70.3
Sønderjylland	194	1229	0.63	1042	84.8
Viborg	181	1636	0.91	1158	70.8
Roskilde	177	1431	0.81	1033	72.2
Ribe	171	1516	0.89	1114	73.5
Bornholm	35	835	2.39	631	75.6
Total	4100	22486	0.55	16690	74.2

Table 2 shows the number of invited and interviewed persons in each sample. The supplementary county sample is the largest, containing almost 48% of all invited persons, whereas the follow-up and the national sample are almost equal in size. In the follow-up sample, the largest part consists of persons invited to the survey in 1994. The response rate in the self-administered questionnaire was 85.5% and is fairly constant across samples.

Table 2. Number of persons invited, interviewed and returning a self-administered questionnaire in the Danish Health and Morbidity Survey 2000

Sample	Invited Number	Interview		Selfadministered questionnaire	
		Number	Response rate (%)	Number	Response rate (%[a/])
National	5,802	4,357	75.1	3,820	87.7
Follow-up	5,912	4,334	73.3	3,662	84.5
1994 survey	5,316	3,884	73.1	3,319	85.0
Age 16-21	486	377	77.6	280	74.3
New Danes	110	73	66.4	63	86.3
Supplementary county	10,772	7,999	74.3	6,796	85.0
Total	22,486	16,690	74.2	14,278	85.5

[a/] Percent of interviewed persons

In the follow-up and the national sample, the county distribution is very similar to the county distribution in Denmark (not shown). Figure 2 shows the county distribution of invited persons in the supplementary county sample compared to population distribution in the Danish population. The counties are sorted as in table 1. As expected, the large counties are undersampled and the small counties are oversampled. However, because the number of persons interviewed exceeds 1,000 and because the response rates vary across counties (table 1), a pattern of increase is less clear than might be expected.

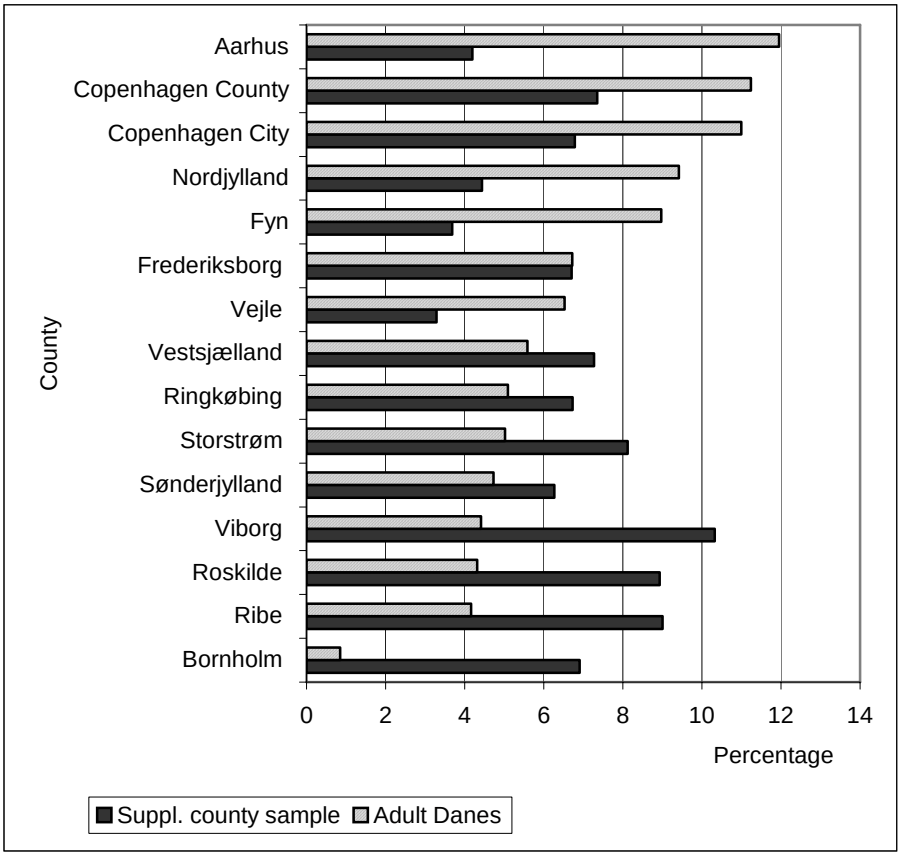


Figure 2 County-specific distribution in the supplementary county sample and among adult Danish citizens 2000, sorted by county population size.

Table 3 shows how well the total sample compares to adult Danes with regard to sex, age-composition and marital status. All information is this table is based on information from the Civil Person Register. These results are available for all adult Danes from Statistics Denmark. As can be seen, the distributions are quite similar for both sexes and in total, thus indicating that the selected sample is representative of the target population.

Table 3. Distribution (5) of age and marital ststus in the 2000 Danish Health and Morbidity Survey (n=22,486) and the adult Danish population (n=4.1 mill.)

		Men		Women		All	
		Total sample	**Popu-lation**	**Total sampl.**	**Popu-lation**	**Total sampl.**	**Popu-lation**
Total		49.0	49.0	51.0	51.0		
Age	16-24	13.3	13.5	12.0	12.5	12.6	13.0
	25-44	35.0	38.1	33.3	35.3	34.1	36.7
	45-66	37.7	34.8	35.1	33.4	36.4	34.1
	67-79	10.3	10.3	13.0	12.4	11.7	11.4
	80+	3.7	3.3	6.7	6.5	5.2	4.9
Marital status*	Unmarried	36.6	37.4	27.9	28.6	32.2	32.9
	Married	51.8	51.1	50.0	49.1	50.9	50.1
	Divorced	7.9	7.9	13.0	12.5	8.4	8.1
	Widowed	3.7	3.5	9.1	9.7	8.5	8.8

***** based on information from Statistics Denmark

3.2. Presentation of results

Table 4 illustrates an example of how the results of the statistical analysis are presented in the final report. The chosen dichotomous indicator is longstanding illness, based on the question 'Do you suffer from any longstanding illness, longstanding after-effect from injury, any disability or other longstanding condition?' (coded as yes / no / don't know). 'Yes', of course, were those responding affirmatively to the question of longstanding illness, while the rest of categories, including missing, were coded as 'No'. Less than 0.1% answered 'don't know' or did not answer at all. Since this question is asked in the interview and in all samples, w_i^{AFU}, normalized to all interviewed, was used as a weight.

The total prevalence and number of persons is shown. Whenever possible, previously published total prevalence from the survey in 1987 and 1994 were

reported along with the prevalence from 2000. As can be seen, the prevalence of longstanding illness increased from 1987 to 2000.

Table 4. Results for longstanding illness in the Danish Health and Morbidity Survey 2000

		Percent	OR		95% confidence interval		Number
Total	1987	32.4					4752
	1994	37.6					4667
	2000	41.1					16690
Men	16-24 years old	28.5	0.48	-	(0.41	- 0.56)	1114
	25-44 years old	33.7	0.61	-	(0.55	- 0.68)	2837
	45-66 years old	44.3	0.96		(0.86	- 1.06)	3063
	67-79 years old	58.7	1.71	+	(1.47	- 1.99)	883
	80 + years old	59.8	1.79	+	(1.40	- 2.29)	291
	All men	40.4					8188
Women	16-24 years old	30.9	0.54	-	(0.46	- 0.62)	1073
	25-44 years old	32.8	0.59	-	(0.53	- 0.65)	2984
	45-66 years old	45.4	1.00				2981
	67-79 years old	59.4	1.76	+	(1.52	- 2.04)	993
	80 + years old	63.5	2.09	+	(1.71	- 2.56)	471
	All women	41.7					8502
ISCED	<10 yeears	53.6	1.56	+	(1.41	- 1.72)	3292
	10 yeears	39.8	1.43	+	(1.24	- 1.65)	1008
	11-12 yeears	46.1	1.40	+	(1.28	- 1.53)	4072
	13-14 yeears	34.3	1.00				4879
	15+ yeears	34.1	0.93		(0.85	- 1.03)	3070
	In school	29.6	-				220
	Other	48.8	-				101
Socio-economic	Self-employed - without subordinate	37.9	1.10		(0.90	- 1.35)	586
group	Self-employed - with subordinate	30.9	0.81	-	(0.66	- 1.00)	557
	Salaried employed I	28.8	0.76	-	(0.64	- 0.90)	1032
	Salaried employed III	31.7	0.92		(0.82	- 1.04)	2611
	Salaried employed I	32.5	1.00				2229
	Skilled worker	29.7	0.84	-	(0.71	- 0.99)	1011
	Unskilled worker	32.9	1.00		(0.86	- 1.15)	1631
	Unemployed	45.1	1.71	+	(1.42	- 2.06)	569
	Pupils/students	30.5	-				1772
	Retirement pensioner	60.3	-				2545
	Disability pensioner	87.6	-				892
	Early retirement allowance	47.4	-				655
	Other	62.4	-				592
Cohabital	Married	40.9	1.00				8692
status	Cohabiting	35.8	1.14	+	(1.03	- 1.26)	2589
	Single (separated, divorced)	52.2	1.55	+	(1.35	- 1.77)	962
	Single (wodowed)	59.6	1.24	+	(1.07	- 1.43)	1264
	Single (unmarried)	35.4	1.27	+	(1.14	- 1.41)	3063
County	Copenhagen City	38.5	0.98		(0.89	- 1.09)	1293
	Copenhagen County	42.0	1.00		(0.90	- 1.10)	1380
	Frederiksborg County	44.9	1.16	+	(1.03	- 1.30)	1100
	Roskilde County	44.7	1.18	+	(1.02	- 1.36)	1033
	Vestsjællands County	37.4	0.83	-	(0.73	- 0.96)	1022
	Storstrøms County	41.0	0.93		(0.81	- 1.07)	1014
	Bornholms County	43.4	1.05		(0.77	- 1.44)	631
	Fyns County	43.2	1.09		(0.98	- 1.21)	1119
	Sønderjyllands County	38.6	0.87	-	(0.76	- 0.99)	1042
	Ribe County	44.5	1.17	+	(1.01	- 1.36)	1114
	Vejle County	39.1	0.92		(0.82	- 1.03)	1032
	Ringkøbing County	37.2	0.86	-	(0.75	- 0.98)	1090
	Aarhus County	40.8	1.03		(0.94	- 1.13)	1474
	Viborg County	40.0	0.95		(0.82	- 1.10)	1158
	Nordjyllands County	42.7	1.07		(0.97	- 1.19)	1188

Shown next is the results for sex and age. The chosen logistic regression model ensures comparability of odds-ratios across all age- and sex-groups. Having an OR but no confidence interval identifies the reference group. Based on the 95% confidence interval, an indication of significance is given. If the lower limit exceeded 1, a '+' is written, if the upper limit is below 1 a '-'.

In Denmark, the educational level has been elevated in the last decades and, therefore, there are more younger persons than older persons with a long education. Thus, the age- and sex-adjusted odds-ratios in groups defined by ISCED in many cases show a different pattern than do the prevalence, especially when the odds-ratio increases (or decreases) across age-groups. In this example, the odds-ratios of having a longstanding illness decreases with increasing level of education. A result like this is interpreted as a social gradient. The two extra categories of ISCED were excluded. The category 'In school' is highly age-dependent in that almost all persons (96%) are 16-24 years old and a model-based description is meaningless for other age groups. The category 'Other' is small and heterogeneous.

For socio-economic status, odds-ratios were estimated only for persons in the labour force. The other groups are either strongly age-dependent, e.g. retirement pensioners (defined as 67 years or older), early retirement allowance (persons aged 60-66 years), or is not judged comparable to other groups, e.g. disability pension (primarily disabled persons aged 45-66 years). In these cases, the prevalence contains the relevant information.

Like ISCED, the age-composition in the groups defined by cohabiting status is inhomogeneous i.e. in general 'widowed' persons tend to be older.

For county, the odds-ratios are deviates from a country average. Inspection of the odds-ratios gives an impression of which counties are above and which are below the country average.

Because the null-hypothsis of equal odds was rejected, a map of equidistant prevalence is displayed (figure 3). As can be seen, the number of counties in each group differs.

Changes over time are shown as prevalence grouped by sex and age (figure 4). The age-groups '67-79' and '80+' being are combined because of the age-groups used in previous surveys. The result of the statistical analysis was not used directly, but a figure (figure 4) was produced based on the final model.

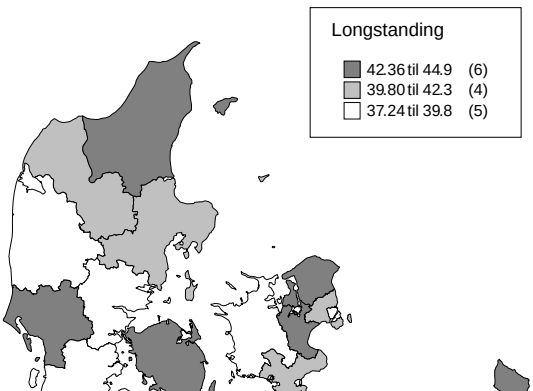


Figure 3. Map of crude grouping of prevalence of longstanding illness in the 15 Danish counties

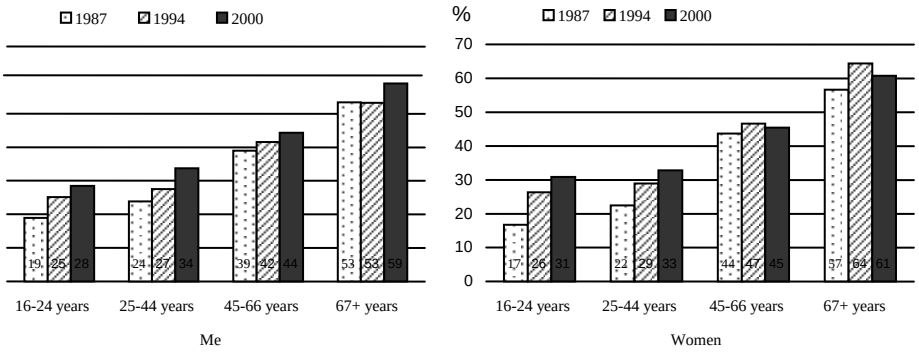


Figure 4. Changes over time of longstanding illness in the Danish Health and Morbidity Survey 2000

4. Discussion

In European surveys [U. S. Department 1999], the sampling methods differ from one country to the other. In some countries, the primary sampling unit is households and persons in the household are interviewed. Likewise, some countries use stratification (but not necessarily on the same units), while others select totally at random. It is likely that these differences depend on other specific

national and local considerations, which, of course, must be taken into consideration when planning a survey.

The idea of incorporating a panel in the survey has been used in Sweden [Statistics Sweden 1996]. Surveys are conducted on a yearly basis including both a panel and a supplement of young aged 16-22.

The chosen design, in fact, implies that the sample size is random. As the goal was to ensure at least 1,000 interviewees in each county, the design depends on the response rate, which was unknown prior to study, in each county. The total number of interviews in each county varied and, thus, no firm stopping rule was employed ('.. at least 1,000 interviews'). Design weights, as in stratified county sampling, were used as we consider the chosen design to be sufficiently close to a design where fixed sampling fractions are defined in each county based on county-specific response rates in the 1994 survey. Also, the sampling fractions are rather small.

We did not use auxiliary information in our post-stratification. The primary reason for this was that in the previous surveys no post-stratification was done and it was essential that the presentations of prevalence from the 1987 and 1994 surveys should be the same as those presented in the previous publications [Rasmussen 1988, Kjøller 1994].

The idea of using odds-ratio comes from the field of epidemiology [Rothman 1998] where they are widely used. There is a great deal of ongoing discussion in this field, concerning the use a multiplicative measure like odds-ratios or absolute measures (differences). Another concern is also whether to use a model-based approach, as is our choice, or to use a more model-free measure such as direct or indirect standardised proportions to obtain age- and sex-adjusted comparisons. We have chosen to use odds-ratios because, when used under conditions that are not too restrictive, e.g. the logistic model, they provide a natural, statistically attractive measure that is easy to interpret. The current description is based on a rather simple version of the logistic regression model enabling s comparison of odds-ratios in groups defined by socio-demographic variables. However, no check of the model is performed and, therefore, the model may be incorrect. Again, the main aim is simple description, rather than thorough analysis. Direct standardisation is more the standard in survey statistics [Arinen 1998].

The SAS system was used to make all calculations both of prevalence and odds-ratios. However, SAS does not take directly into account the design, while the SUDAAN software package [Shah 1997] has facilities to this end. In this study, persons were sampled at random and the sampling fractions in each county were rather small (table 1). All but one (2.4%) was below 1%. As we consider this design as stratified on county level, it is expected that the difference between SAS and SUDAAN is of limited practical importance. A small sensitivity analysis based on 4 indicators showed that the standard errors for the log-odds of the socio-demographic variables was approximately 5% lower in SAS than in SUDAAN. It should be mentioned that standard errors of prevalence were, as

expected, higher in SAS than in SUDAAN. The same weights were used in SAS and SUDAAN and, hence, the odds-ratios obtained are exactly the same. In SUDAAN, we used Taylor's series expansion to calculate the standard errors.

Thus, we have presented a complex survey arising due to multiple purposes within the survey. The design and the weighting system used have been described in detail and the statistical analysis underlying the presentation of results has been documented. An example of how the results are presented has been given. The choices made have been discussed and the scope of health interview surveys, of which this survey is an example, has been outlined.

Acknowledgment

The authors would like to thank the referees for many helpful comments especially concerning the structure of the paper.

REFERENCES

ARINEN S, HÄKKINEN U, KLAUKKA T, KLAVUS J, LEHTONEN R, ARO S: Health and the Use of Health Services in Finland. Main findings of the Finnish Health Care Survey 1995/96 and changes from 1987. Gummerus Kiriapaino Oy; 1998.

HOSMER DW, LEMESHOW S: Applied Logistic Regression. Wiley, New York. 2000.

HUPKENS CLH, van der BERG J, van der ZEE J: National health interview surveys in Europe: an overview. Health Policy; 47(2): 145-68. 1999.

KJØLLER M, RASMUSSEN NK, KEIDING L, PETERSEN HC, NIELSEN GA: Health and Morbidity in Denmark 1994 – and the development since 1987. Copenhagen: National Institute of Public Health, 1995 (in Danish).

KJØLLER M, RASMUSSEN NK (ed): Health and Morbidity in Denmark 2000. Copenhagen: National Institute of Public Health, 2002 (in Danish).

RASMUSSEN NK, GROTH MV, BREDKJÆR SR: Heath and Morbidity in Denmark 1987. Copenhagen; National Institute of Public Health, 1988 (in Danish).

ROTHMAN JK, Greenland S: Modern epidemiology (second edition).

Lippincott-Raven; 1998.

SAS/STAT User Guide, Version 8. SAS Institute Inc., Cary, NC. 1999.

SHAH BV, BARWELL BG and BIELER GS: SUDAAN User's Manual, Release 7.5. Research Triangle Park, NC; Research Triangle Institute. 1997.

STATISTICS SWEDEN (1995): Levnadsförhållanden. Appendix 15: Teknisk rapport 1990-1993. Statistics Sweden, 1995 (in Swedish).

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVIES (1999): International Health Data Reference Guide, 1999. National Center for Health Statistics, Hyatsville, MD; USA.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 943–968

ESTIMATING TOTALS IN SOME U.K. BUSINESS SURVEYS

Dan Hedlin[1]

ABSTRACT

The paper discusses estimation of the total for some study variables in two business surveys conducted by the U.K. Office for National Statistics. The paper is mainly focused on design-based methods. We ask what the desirable properties of an estimator are and explore several point estimators in a simulation study.

Key words: model groups, local regression estimator, robust regression estimator, mixture model estimator.

1. Introduction

Since business data are skewed, outlier prone and often contain a large proportion of zeroes, it is not obvious that traditional methods of using auxiliary data, e.g. ratio and regression estimation, have the properties they often are believed to have, such as being virtually free from bias and have competitive variance. We study whether this is always true for real populations and if the total can be estimated more accurately *and/or* more robustly by either robustifying these instances of the generalised regression (GREG) estimator (Särndal, Swensson, and Wretman 1992), or by relying more explicitly on a model.

Most surveys at the U.K. Office for National Statistics (ONS) are multipurpose with customers who will use the statistics in different ways. The estimated totals for business surveys are, however, particularly important as they are input to the National Accounts.

What properties of an estimator of the total are vital? One could think of, e.g., small variance, negligible bias, good confidence intervals or minimum risk of obtaining estimates with large error; or versatility or ease of implementation. We conduct a simulation study in which several GREG estimators are compared with a not widely used local regression estimator and a robust regression estimator that

[1] Department of Social Statistics, University of Southampton, Southampton SO17 1BJ, U.K.

is novel in a design-based context. The former is similar to the GREG but has the ability to accommodate local departures from the underlying linear model.

For many estimators there is a choice of model groups to be made. For example, a ratio model can be fitted within strata (the separate ratio estimator) or across strata (the combined ratio estimator), see Cochran (1977). There is little research on how to choose model groups. Silva and Skinner (1997) minimise the mean squared error to find the optimal set of auxiliary variables and thereby also model groups. We simulate three types of model group partitions and compute five criteria for each combination of estimator and type of model group partition. Two of the criteria are rather non-traditional.

Many of the business surveys at the ONS use a stratified simple random sampling design with four size strata within industry, three of which are genuine sampling strata and the one with the largest units is a completely enumerated (CE) stratum. There are two interval scaled variables on the frame: register employment and turnover. Industries are important domains of study.

Size strata (employment sizebands within a domain)	Strategy
4	A completely enumerated stratum + the separate ratio estimator to account for nonresponse
3 2 1	} Genuine sampling strata + combined ratio estimation

Figure 1. Current ONS sampling and estimation strategy in an industry

There are typically four size strata (sizebands) within a domain, see Figure 1; sizeband 1 ($20 \leq \text{employment} \leq 49$) and sizeband 4 (employment ≥ 300) comprise the smallest and the largest units respectively. Sizeband 4 is completely enumerated, although some nonresponse occurs. We will, however, assume full response and ignore measurement errors and incomplete coverage of the target population.

In Section 2 the model groups and estimators used in the simulation study are defined, whose results are reported in Section 3. The paper ends with a discussion in Section 4.

2. Estimators

2.1. Aim

The aim here is to estimate the total $t_y = \sum_U y_k$ of a study variable $\mathbf{y}' = (y_1, y_2, \ldots, y_N)$ on a population U with unit labels $\{1,2,\ldots, N\}$. It is

assumed that there is an auxiliary variable $\mathbf{x}' = (x_1, x_2, \ldots, x_N)$, with x_k known for each element in U. A sample s of size n is taken and (x_k, y_k) is observed for all units k in the sample. Let stratum quantities and sets be indexed by h. For example, N_h and s_h refer to stratum size and the sample that is taken from stratum h. The populations of interest are industries. We assume that all units are correctly classified to industries before the sample is drawn. The terms domain (industry) and population will be used interchangeably.

2.2. Model groups

For many estimators there is a choice of population partitioning that defines the G '***groups***' (subsets) in which the model is to be fitted, $U = U_1 \cup \ldots \cup U_g \cup \ldots \cup U_G$ (Särndal et al. 1992, Sec 7.5). We study three types of model group partition:

a) groups coincide with strata;
b) one group consists of all genuine sampling strata and another group of the CE stratum within the industry (see Figure 1);
c) all strata within an industry constitute one group.

Case *b*, '**ONS model groups**', is the type of partition into model groups ONS use for many business surveys. Cases *a* and *c* are labelled '**within strata**' and '**over all strata**'.

2.3. Point estimators

Many estimators used in practice are of the form

$$\hat{t} = \sum_{k \in U} \hat{y}_k + \sum_{j \in s} \tilde{\omega}_j (y_j - \hat{y}_j), \qquad (1)$$

a 'model-based' or 'synthetic' term plus a 'bias adjustment' or 'correction' term, with predicted values $\hat{y}_k = \boldsymbol{\omega}_k' \mathbf{y}_s$, $\mathbf{y}_s' = (y_1, y_2, \ldots, y_n)$, and weight vector $\boldsymbol{\omega}_k$ and scalar $\tilde{\omega}_k$, neither dependent on $\mathbf{y}$. The weights may be sample dependent. We can refer to an estimator that can be written on the form (1) as an 'projective bias adjusted estimator' ('projective' because the predicted values are projected to non-sample units or all population units). It can be shown that this estimator is linear, i.e., it can be written as a sample sum of the products of some weights, not dependent on $\mathbf{y}$, and the y_k. This property is highly desirable from a national statistical institute's point of view. The main reason is practical: for example, the weights can be thought of as 'grossing factors', stored in one column in a file and be applied in a simple way to all study variables without recomputation. Also, estimates of a linear estimator are internally consistent in the sense that if $\hat{t}_i$ is an estimator for a variable i, then $\hat{t}_1 + \hat{t}_2 = \hat{t}_{1+2}$ for the sum of the variables.

Theoretical arguments do not abound, but one reason put forward by Sugden and Smith (2002), is that if the parameter is a single sum of a function of the population units (most parameters of practical interest are) then the estimator must have the same form if it is going to reduce to the parameter when $n = N$.

What the weight vectors are and how $\hat{\mathbf{y}}_s$ is computed may depend on the sampling design and the way the model is fitted. For example, the predicted values may be obtained with a least squares fit or with a model-assisted approach involving the inverse inclusion probabilities as weights. If $\boldsymbol{\omega}_k$ has zeroes in elements 'far' from element k then only elements in the vicinity of k in $\mathbf{y}_s$ will contribute to the predicted value $\hat{y}_k$. Thus there is a trade-off between using as many observations as possible to predict $\hat{y}_k$ and not letting possibly less relevant observations far away from k play a role. Also, there is a decision to make about the exact impact of units in the vicinity of k and that of those further away.

For some estimators the bias adjustment term always is zero (e.g. the ratio estimator), for some other estimators it will take whatever value to achieve some overall property. For example, a regression estimator corresponding to a heteroscedastic regression model for y on a variable x with heteroscedasticity proportional to $x^{1.5}$ or x^2 is asymptotically design-unbiased only if the bias adjustment is allowed to be unbounded. Hedlin, Falvey, Chambers, and Kokic (2001) show that this can lead to extremely poor performance for these estimators. In a model-based setting the bias adjustment term can explicitly be regarded as an estimate of the bias due to model misspecification (Chambers, Dorfman, and Wehrly 1993). If a model M^*, $y_k = m(\mathbf{x}_k) + \varepsilon_k$, say, is correct and $\hat{t}$ is based on another, working model M (perhaps a simpler model than M^*) then the bias is $\sum_{k \in U-s} v_k$, where v_k denotes the non-observed residuals $\hat{y}_k - m(\mathbf{x}_k)$ for non-sample points. $E_M(v_k)$ can be estimated from the observed residuals r_j: $\sum_{k \in U-s} \hat{v}_k = \sum_{k \in U-s} \sum_{j \in s} \ddot{\omega}_{jk} r_j = \sum_{j \in s} \ddot{\omega}_{j\bullet} r_j$ with some appropriate weights. Hence this term can be viewed as an estimate of the bias due to model misspecification and a special case of the bias adjustment term in (1). In a design-based framework such as GREG estimation, the second term of (1) may be $\sum_{j \in s} \pi_j^{-1}(y_j - \hat{y}_j)$; this term estimates $-\sum_{k \in U} v_k = t_y - \sum_{k \in U} \hat{y}_k$.

In general, there is an interplay between the choice of $\boldsymbol{\omega}_k$ and $\tilde{\omega}_k$. The bias adjustment term should normally be far smaller than the 'model-based' term. If not, there is an indication of model misspecification or a dysfunctional relationship between the structure of the data and what you do with them. As

shown by Hedlin et al. (2001), the ratio of the bias adjustment term to $\sum_U \hat{y}_k$ is an important diagnostic for some GREG estimators, where a large value indicates model problems. However, not all estimators offer flexibility in the choice of weights $\boldsymbol{\omega}_k$ and $\tilde{\omega}_k$. For those estimators, once the estimator is chosen one has to accept the weights that the estimator prescribes.

Winsorisation is one way of curbing the influence of outliers that is not included in this study. Winsorisation is a value-modification strategy where the value of a sampled unit is adjusted downwards if it is larger than a predefined cut-off (Kokic and Bell 1994). Value modification could be viewed as artificial and hence it may run the risk of not gaining public acceptance. Furthermore, the main argument for Winsorisation is that of minimum mean squared error, even if it comes at the expense of a large bias. Minimum MSE may be strong argument for some surveys but less so for others. Many other outlier-robust estimators have been proposed, in particular model-based ones. Overviews include Chambers and Kokic (1993) and Valliant, Dorfman, and Royall (2000, Ch 11).

Below follows a detailed description of the estimators used in the simulations.

2.3.1. The Horvitz-Thompson estimator

Let

$$\hat{t}_{yg\pi} = \sum_{s_g} w_k y_k \tag{2}$$

be the expansion estimator for the group total $t_{yg} = \sum_U y_k I(k \in g)$, where $I(k \in g) = 1$ if the sampled unit *k* belongs to group *g*, and 0 otherwise. Here s_g is the part of the sample that falls in group *g* and $w_k = \pi_k^{-1}$ is the inverse sampling weight for unit *k*. Let $\hat{t}_{y\pi}$ be the expansion estimator '***HT***' for the total t_y in *U* (i.e., the sum of the group estimates). The HT estimator for stratified simple random sampling is a special case of (1).

2.3.2. GREG estimators

The ratio estimator '***Rat***' for some set of model groups is (Särndal et al. 1992, Sec. 7.7):

$$\hat{t}_{yrat} = \sum_{g=1}^{G} t_{xg} \frac{\hat{t}_{yg\pi}}{\hat{t}_{xg\pi}}, \tag{3}$$

where *x* is an auxiliary scalar. Rat combined with the ONS type of model group is the estimator the ONS uses for many business surveys.

The GREG estimator can be written (Särndal et al. 1992, Ch. 6) as

$$\hat{t}_{yreg} = \sum_U \hat{y}_k + \sum_s w_k \left(y_k - \hat{y}_k\right), \tag{4}$$

where $\hat{y}_k = \mathbf{x}'_{kg}\hat{\mathbf{B}}_g$, $\hat{\mathbf{B}}_g = \left(\mathbf{X}'_{sg}\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_{sg}\right)^{-1}\mathbf{X}'_{sg}\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{y}_{sg}$, and $\mathbf{X}_{sg}$ is a p x n matrix with $\mathbf{x}'_{kg}$ in the *kth* row (two special cases to be given shortly), and $\boldsymbol{\Pi}_s$ and $\boldsymbol{\Sigma}_s$ are diagonal matrixes with π_k and the residual variance σ_k^2 in position (k, k), respectively. The data are assumed to follow a superpopulation model M for which $E_M\left(Y_k\right) = \mathbf{x}'_{kg}\boldsymbol{\beta}_g$ and $V_M\left(Y_k\right) = \sigma_k^2$, k = 1, 2, …, N, where the moments are taken over the model. The Rat estimator $\hat{t}_{yrat}$ is a special case of (4) with $\mathbf{x}'_{kg} = I\left(k \in g\right)x_k$ and $V_M\left(Y_k\right) = \sigma^2 x_k$, k = 1, 2, …, N. The '***Reg***' estimator is another special case with $\mathbf{x}'_{kg} = I\left(k \in g\right)\begin{pmatrix}1 & x_k\end{pmatrix}$. For '***Reg/1.0***' we take $V_M\left(Y_k\right) = \sigma^2 x_k$, and for '***Reg/1.5***' $V_M\left(Y_k\right) = \sigma^2 x_k^{1.5}$.

Reg/1.5 is the linear estimator that is associated with the model M that gives the best fit to the data used in simulations reported below. Hence, we would expect good performance for Reg/1.5.

GREGs are projective bias adjustment estimators. There is an intriguing, qualitative difference between Reg/1.5 and Reg/1.0. Under the model associated with Reg/1.0, (4) reduces to $\hat{t}_{yreg} = \sum_U \hat{y}_k$ (Särndal et al., 1992, p. 231).

2.3.3. Local and robust regression estimators

To make the GREG estimators more robust to outliers and a nonlinear relationship between the study and auxiliary variable, the $\hat{y}_k$ in (4) can be replaced with some robust prediction. Breidt and Opsomer (2000) use a local polynomial regression estimator weighted with w_k to produce predictions $\hat{m}_k$ that in many cases will be close to $\hat{y}_k$. The estimator, here referred to as '***Local***', is

$$\hat{t}_{yloc} = \sum_U \hat{m}_k + \sum_s w_k \left(y_k - \hat{m}_k\right). \tag{5}$$

Chambers et al. (1993) suggested a similar but model-based estimator. A somewhat less general estimator than Breidt's and Opsomer's is

$$\hat{m}_k = \mathbf{e}'_{(2)1}\left(\mathbf{D}'_k\mathbf{W}_k\mathbf{D}_k\right)^{-1}\mathbf{D}'_k\mathbf{W}_k\mathbf{y}_s,\ k = 1, 2, \ldots, N, \tag{6}$$

where $\mathbf{e}'_{(d)j}$ is a d-vector with 1 in the *jth* position and 0s otherwise, $\mathbf{D}_k$, k = 1,

2, …, *N,* are n x 2 matrices, each with $\begin{bmatrix} 1 & (x_j - x_k) \end{bmatrix}$ in the *jth* row, $j = 1, 2, \ldots, n$; $\mathbf{W}_k$, for $k = 1, 2, \ldots, N$, are n x n diagonal matrices with $w_k b_k^{-1} K\left[(x_j - x_k) b_k^{-1}\right]$ in cell (j, j) with $K(\cdot)$ and b_k being the kernel function and the bandwidth, respectively. For a fixed bandwidth, Breidt and Opsomer prove that the sample weights in $\mathbf{W}_k$ and in (5) make $\hat{t}_{yloc}$ asymptotically design-unbiased. Their estimator has several other desirable theoretical properties. Like them we use the Epanechikov kernel

$$K(u_{jk}) = \max\left[0, \frac{3}{4}(1 - u_{jk}^2)\right]. \tag{7}$$

For a fixed bandwidth, the minimum bandwidth would have to be longer than the longest distance between two consecutive x-values, which for skewed populations would prohibit truly local regression. Therefore, we use two types of variable bandwidth, first

$$b_k^{(s)} = x_{k+20} - x_{k-20}, \tag{8}$$

where $x_{k\text{-}20}$ and x_{k+20} are units in the sample file, sorted by x_k in ascending order. Note that $K(u_{jk}) = 0$ if $u_{jk} = (x_j - x_k) / b_k^{(s)} \geq 1$. Hence the kernel defines a window around unit k outside which units will not contribute to the prediction of $\hat{m}_k$. The window slides across stratum boundaries including the CE stratum (sizeband 4). Note that $\hat{m}_k$ will cancel out in the CE-stratum in (5), although the 20 smallest units in this stratum will be used in prediction of units in sizeband 3. If k is so small that $x_{k\text{-}20}$ does not exist, $x_{k\text{-}20}$ is taken as the minimum x-value and similarly for x_{k+20}. No adjustment has been made for these boundary effects. For the other type of bandwidth,

$$b_k^{(f)} = x_{k+40} - x_{k-40}, \tag{9}$$

x_{k+40} and $x_{k\text{-}40}$ are taken from the frame sorted by x_k in ascending order. The number of sample units in the window will vary with the π_k: for parts of the frame with small sample fractions the local regression fit will tend to be more wiggly than in more densely sampled areas. It seems reasonable that a point in a lightly sampled stratum should be given more influence. Care must be taken so that $\mathbf{D}_k' \mathbf{W}_k \mathbf{D}_k$ is not singular. The local regression estimators with bandwidths (8) and (9) are labelled ***Local/s20*** and ***Local/f40***, respectively.

The prediction (6) can be rewritten as

$$\hat{m}_k = \bar{y}_{loc} + \left(x_k - \bar{x}_{loc}\right)\frac{\sum_{j\in s} q_{jk}\left(x_j - \bar{x}_{loc}\right) y_j}{\sum_{j\in s} q_{jk}\left(x_j - \bar{x}_{loc}\right)^2} \qquad (10)$$

where the q_{jk} are diagonal elements in the $\mathbf{W_k}$, $\bar{y}_{loc} = \sum_{j\in s} q_{jk} y_j \left(\sum_{j\in s} q_{jk}\right)^{-1}$, and similarly for $\bar{x}_{loc}$. Formulation (10) shows that the local linear prediction is $\bar{y}_{loc}$ (which is what would have been obtained with local constant prediction without the *x*-variable) plus a term that counteracts effects stemming from the local slope and boundary effects.

Let *j* and *k* index sample and population units, respectively. Note that (5) can be written

$$\begin{aligned}\hat{t}_{yloc} &= \sum\nolimits_s w_j y_j + \sum\nolimits_U \left[1 - I\left(k \in s\right) w_k\right] \hat{m}_k \\ &= \sum_{j\in s} w_j y_j + \sum_{j\in s}\left\{\sum_{k\in U}\left[1 - I\left(k \in s\right) w_k\right] \mathbf{e}'_{(2)1}\left(\mathbf{D}'_k \mathbf{W}_k \mathbf{D}_k\right)^{-1} \mathbf{D}'_k \mathbf{W}_k \mathbf{e}_{(n)j}\right\} y_j \qquad (11)\end{aligned}$$

that is, $\hat{t}_{yloc} = \sum\nolimits_s w_{loc,js} y_j$ is a linear estimator with weights

$$w_{loc,js} = w_j + \sum_{k\in U}\left[1 - I\left(k \in s\right) w_k\right] \mathbf{e}'_{(2)1}\left(\mathbf{D}'_k \mathbf{W}_k \mathbf{D}_k\right)^{-1} \mathbf{D}'_k \mathbf{W}_k \, \mathbf{e}_{(n)j}. \qquad (12)$$

The subscript *s* reminds us that the weights are sample dependent. To continue the analogy with the GREG estimator, the local regression estimator weights can be partitioned into sampling weights w_j and 'local *g*-weights'

$$g_{loc,js} = 1 + \frac{1}{w_j}\left\{\sum\nolimits_U \left[1 - I\left(k \in s\right) w_k\right] \mathbf{e}'_{(2)1}\left(\mathbf{D}'_k \mathbf{W}_k \mathbf{D}_k\right)^{-1} \mathbf{D}'_k \mathbf{W}_k\right\} \mathbf{e}_{(n)j} \qquad (13)$$

The Local/f40, Local/s20 and the HT estimators are the only estimators in this paper that do not depend on the partitioning of the population into model groups. The Local estimators are of the projective bias adjustment form (1). They are flexible in that for a long bandwidth they will be similar to the GREG, and for a shorter bandwidth they will capture local model departures. Different kernels will give different distributions of weights **ω** within the window. The main difference between our and Briedt's and Opsomer's versions is our use of variable bandwidths.

Another estimator here called ***RobReg/f40***, was inspired by Chambers and Dunstan (1986) and Kuk and Welsh (2001). One difference is that we take a

design-based approach. In (6), $\mathbf{y}_s$ is replaced with $\tilde{\mathbf{y}}_k = \left(\mathbf{x}_k\tilde{\boldsymbol{\beta}}_g + x_k^{3/4}\tilde{r}_1, ..., \mathbf{x}_k\tilde{\boldsymbol{\beta}}_g + x_k^{3/4}\tilde{r}_n\right)'$, where the tilde indicates a robust fit obtained with bounded-influence estimation (to be specified shortly), to produce a smoothed value $\hat{m}_k^*$. The advantage of projecting $\tilde{\mathbf{r}}$ to each frame unit k is to allow for an asymmetric distribution of the residuals. The value $\hat{m}_k^*$ may well be some distance away from $\tilde{\mathbf{y}}_k$. Hence smoothing is done in two dimensions, first horizontally through the robust regression, then vertically through the smoothing of each $\tilde{\mathbf{y}}_k$ separately. It is conjectured that RobReg is approximately design-unbiased.

The bounded-influence method utilises the $DFFITS_k$ of each observation k, which is a well known measure of how much the prediction for this observation's x-value would change in terms of standard deviations of the predicted value if the regression line is refitted without observation k. Welsch (1980) suggests the use of the inverse $DFFITS_k$ as regression weights, a method analysed by Ryan (1997, Ch. 11). Belsley, Kuh and Welsch (1980) suggest as a rule of thumb for univariate regression that observations with larger absolute value of $DFFITS_k$ than $2n^{-0.5}$, n being the number of observations, should get special attention. The regression weights proposed by Welsch are

$$\delta_k = \begin{cases} 1 \text{ if } \left|DFFITS_k\right| \le 2n^{-0.5} \\ 2n^{-0.5}\left|DFFITS_k\right|^{-1} \text{ if } \left|DFFITS_k\right| > 2n^{-0.5} \end{cases} \tag{14}$$

The regression parameters are estimated with weighted least squares with the weights $\delta_k x_k^{-3/4}$. The residuals are

$$\tilde{r}_j = \frac{y_j - \mathbf{x}_{kg}\tilde{\boldsymbol{\beta}}_g}{x_k^{3/4}} \tag{15}$$

RobReg is an estimator robust to outliers. However, it is not of form (1). It is not linear and it has not the internal consistency property. Theoretical properties such as bias will be developed elsewhere.

2.3.4. Mixture model estimators

The Karlberg (2000) estimator can be seen as a transformation–retransformation estimator. It is based on a mixture model. Let Z_k be the logarithm of the study variable $Y_k > 0$. Assume that $y_1, y_2, ..., y_N$ are realisations of the random variables $Y_1, Y_2, ..., Y_N$, and, conditional on the auxiliary variable,

$E_{\ddot{M}}\left(Z_k \middle| Y_k > 0\right) = \mu_g = \mathbf{x}'_{kg}\boldsymbol{\beta}_g$, $V_{\ddot{M}}\left(Z_k \middle| Y_k > 0\right) = \sigma_g^2$ where $\mathbf{x}'_{kg} = \mathbf{1}(k \in g)(1 \quad x_{2k})$, with x_{2k} being the logarithm of the auxiliary variable, provided that $x_{2k} > 0$. The parameter $\boldsymbol{\beta}_g$ is estimated through OLS regression applied to the logtransformed data. The model $\ddot{M}$ differs from that of Karlberg (2000) in that we allow for different model groups but not heteroscedasticity. Not to burden the notation we will suppress subscript *g* from now. Let **X** be the matrix with $\mathbf{x}'_{kg}$ in the *kth* row, and let subscript *s* indicate the corresponding sample entity. To estimate the total of the nonsampled units on the original scale, the sum of the back-transformed predicted values of the study variable are multiplied by a bias correction factor. Let a_{kk} be the diagonal elements in a matrix $\mathbf{X}\left(\mathbf{X}'_{s+}\mathbf{X}_{s+}\right)^{-1}\mathbf{X}'$, which is rather similar to the 'hat matrix', with s+ indicating that the matrix is restricted to positive sample values. If $\hat{Z}_k$ is the predicted value for unit *k* on the logscale, i.e. $\hat{Z}_k = \mathbf{x}'_k\hat{\boldsymbol{\beta}}$, then $\exp\left(\hat{Z}_k\right)$, whose expected value is $\exp\left(\mu + a_{kk}\sigma^2/2\right)$, is a biased estimate of the value Y_k on the original scale. Based on the additional assumption that Y_k follows a lognormal distribution with mean and variance given by $\ddot{M}$, so that $E_{\ddot{M}}\left(Y_k\right) = \exp\left(\mu + \sigma^2/2\right)$, Karlberg derives a approximately model-unbiased predictor:

$$\hat{Y}_k = \exp\left(\hat{Z}_k\right)\exp\left[\frac{\hat{\sigma}^2}{2}\left(1 - a_{kk}\right) - \frac{\hat{\sigma}^4}{4n_+}\right]. \tag{16}$$

where n_+ is the number of positive elements in the model group and

$$\hat{\sigma}^2 = \frac{\mathbf{Z}'_{s+}\mathbf{Z}_{s+} - \hat{\beta}'\mathbf{X}'_{s+}\mathbf{X}_{s+}\hat{\beta}}{n_+ - 2} = \sum_{k=1}^{n_+} e_k^2 \Big/ \left(n_+ - 2\right), \tag{17}$$

with e_k being the residuals on the logscale. If $n_+ \leq 2$, then the denominator of (17) is set to 1. The ***Logn/pr*** estimate of a total for a model group *g* is

$$\hat{T} = \sum_{h=1}^{H}\sum_{k=1}^{N_g - n_g}\hat{p}_h\hat{Y}_k + \sum_{k=1}^{n_g} Y_k \tag{18}$$

where $\hat{p}_h$ is the sample proportion of positives in sizeband *h*. Alternatively, a logistic model is fitted within each sizeband to obtain an estimated probability $\hat{p}_h(x)$ for a unit with a certain *x*-value to have a positive *y*-value. This estimator is labelled ***Logn/log***.

In the simulations it often happened that the two groups defined by whether the study variable is zero or not were completely separated. For example, if all *x*-values for zero study variable values are smaller than those of the positive study variable values, then the groups are completely separated and no ML estimates of the parameters of the logistic model exist. In this case, $\hat{p}_h(x)$ was set to one for *x*-values greater than the average of the largest and smallest of the sample *x*-values on either side of the separation point, and zero otherwise. For the rather more unlikely contingency that the groups were completely separated apart from one shared sample x-value ('quasi-complete separation'), $\hat{p}_h(x)$ was set to ½ for the shared point. If the sample x-values overlap the ML estimates exist and are unique. Overlap, complete and quasi-complete separation partition the space of data configurations (Albert and Anderson 1984).

The mixture model estimators are sensitive to errors in $\hat{\sigma}^2$. Therefore, ***RLogn/pr*** is obtained by using a robust estimate of the variance, $\hat{\sigma}_R^2$. The beta coefficient $\hat{\beta}_R$ was computed through a regression relationship within model groups of $\log(y_k)$ on $\log(x_k)$, with homoscedastic errors and weights (14). The estimate $\hat{\sigma}_R^2$ was taken as 1.4826 times the median absolute deviation of the residuals $y_k - \hat{\beta}_R x_k$ from their median. The constant 1.4826 is chosen so as to make $\hat{\sigma}_R^2$ consistent if the errors were standard normal.

The mixture model estimators are attractive in their relative simplicity, but they are not in general design unbiased. They cannot be written on the form (1). Transforming to log scale makes many business survey datasets nicely linear, apart from the zero-valued observations. The flipside is the need to estimate the potentially influential parameter σ^2 and, as a consequence of the lognormal model assumption, the need to estimate the propensity for a unit to have a zero value. The partition of the sample data into positives and zeroes makes the effective sample data set smaller.

2.4. Variance estimators

Although this paper focuses on point estimation, we have computed coverage probabilities and hence variance estimates. For a general 'g-weighted' variance estimator for GREGs see Särndal et al. (1992, Ch 6). It can be shown that a g-weighted variance estimator for $\hat{t}_{yrat}$ for the three types of model group combined with stratified simple random sampling (STSI) is

$$\hat{V}_{STSI}\left(\hat{t}_{rat}\right) = \sum_{h=1}^{H}\left(\frac{t_{xg}}{\hat{t}_{xg\pi}}\right)^2\left[N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\sum_{s_h}\left(e_k-\bar{e}_h\right)^2\right] \qquad (19)$$

where $e_k = y_k - x_k' \hat{B}_{ratg}$ with $\hat{B}_{ratg} = \dfrac{\hat{t}_{yg\pi}}{\hat{t}_{xg\pi}}$. Here the g-weights are $g_{ks} = t_{xg} / \hat{t}_{xg\pi}$. For example, for ONS model groups, (19) is

$$\hat{V}_{STSI}\left(\hat{t}_{rat}\right) = \left(\frac{t_{x1}}{\hat{t}_{x1\pi}}\right)^2 \sum_{h=1}^{3}\left[N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{1}{n_h - 1}\sum\nolimits_{s_h}\left(e_k - \bar{e}_h\right)^2 \right], \tag{20}$$

where the totals in group $g = 1$ (all genuine sampling strata $h = 1$, 2, and 3) are

$$\hat{t}_{x1\pi} = \sum_{h=1}^{3}\sum\nolimits_{s_h} w_k x_k \text{ and } t_{x1} = \sum_{h=1}^{3}\sum\nolimits_{U_h} x_k \ .$$

It can also be shown that the g-weighted variance estimator for the group regression model is

$$\hat{V}_{STSI}\left(\hat{t}_{reg}\right) = \sum_{h=1}^{H}\left[N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{1}{n_h - 1}\sum\nolimits_{s_h} g_{ks}^2\left(e_k - \bar{e}_h\right)^2 \right], \tag{21}$$

where $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_g$ with $\hat{\mathbf{B}}_g$ defined in Sec 2.3.2, and

$$g_{ks} = 1 + \left(\mathbf{t}_{xg} - \hat{\mathbf{t}}_{xg\pi}\right)' \left(\sum\nolimits_s w_j \mathbf{x}_{jg}\, \mathbf{x}_{jg}' / \sigma_j^2\right)^{-1} \left(\mathbf{x}_{kg} / \sigma_k^2\right). \tag{22}$$

The g-weights for regression models are studied in great detail by Hedlin et al. (2001).

The variance estimator used here for Local is

$$\hat{V}_{STSI}\left(\hat{t}_{yloc}\right) = \sum_{h=1}^{H} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{S_h^2}{n_h}, \tag{23}$$

where $S_h^2 = \dfrac{\sum_{s_h}\left(\gamma_k - \bar{\gamma}_h\right)^2}{n_h - 1}$, $\gamma_k = y_k - \hat{m}_k$, and $\bar{\gamma}_h = n^{-1}\sum_{s_h}\gamma_k$.

Breidt and Opsomer (2000) show that (23) is for a fixed bandwidth a consistent estimator of an approximate variance

$$AV\left(\hat{t}_{yloc}\right) = \sum\sum\nolimits_U \Delta_{kl}\, \Gamma_k \Gamma_l / \pi_k \pi_l \ , \tag{24}$$

where $\Gamma_k = y_k - m_k$, m_k being the smoothed values one would get with (6) based on the whole population, and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ with π_{kl} being the probability that both units k and l are included in the sample. The expression (24) has the same form as the usual approximate variance of the GREG, see Särndal et al. (1992, Ch 6). The local g-weights (13) could be inserted into (23). The estimator (23) was used for RobReg as well with $y_k - \hat{m}_k^*$ replacing γ_k.

We have not computed variance estimates for the mixture model estimators.

3. Simulations based on MIDSS and CAPEX data

Some domains of the Quarterly Survey of Capital Expenditure (CAPEX) and the Monthly Inquiry for the Distribution And Services Sector (MIDSS), both conducted by the ONS, provided data for a simulation study. The sampling design for both surveys is as shown in Figure 1. The study variable is turnover for the MIDSS. Here net capital expenditure was used as the CAPEX study variable. For the purposes of this study, the auxiliary variable for both the MIDSS and the CAPEX was turnover, which is the variable that correlates most strongly with either of the study variables.

Figures 2 to 6 show scatter plots of three MIDSS and two CAPEX domains on logscale. For confidentiality reasons the scales of the axes are suppressed. Note that the CAPEX domains U and V are very different from the MIDSS domains A, B and C. Note in particular that the largest value of the auxiliary variable in domain V is in sizeband 3, i.e. a sampled sizeband. Domain V is the domain studied by Hedlin et al. (2001). The proportion zero values for the study variable is rather small for the MIDSS. For the CAPEX it is about 40% and 20% for domains U and V respectively.

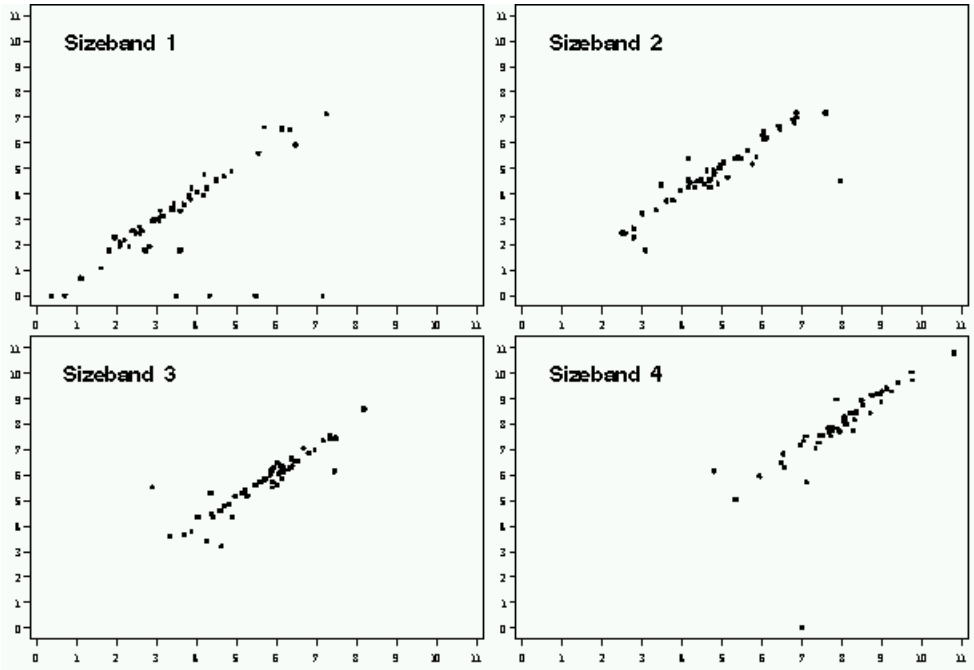


Figure 2. MIDSS, domain A. Log of the study variable against log of the auxiliary variable, with unity added to both variables

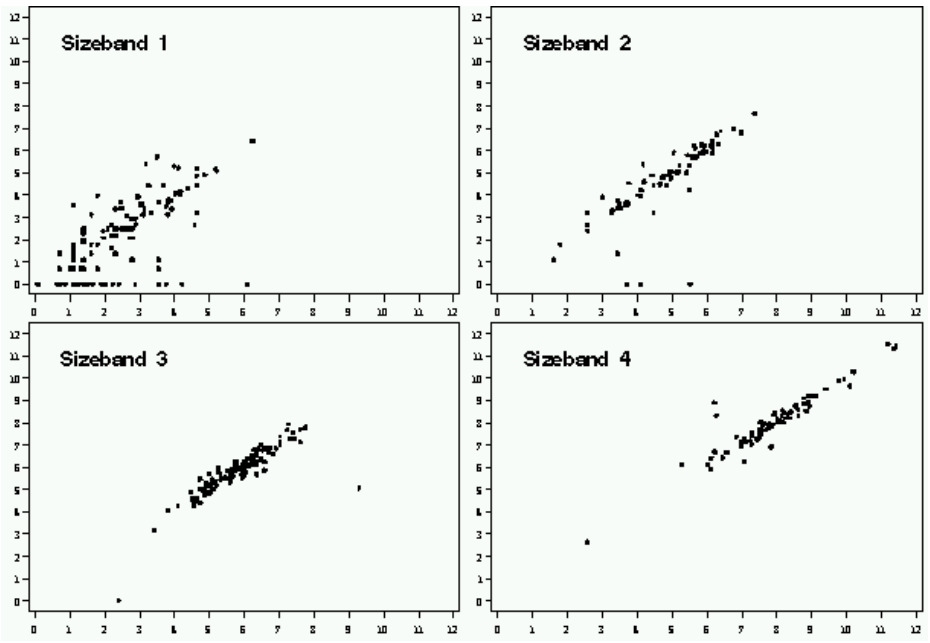


Figure 3. MIDSS, domain B. Log of the study variable against log of the auxiliary variable, with unity added to both variables

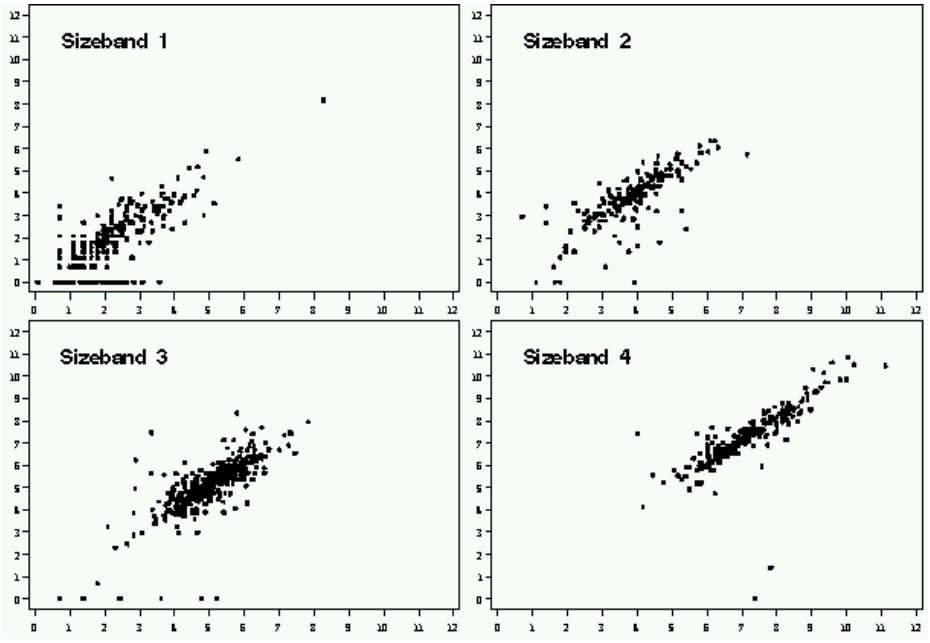


Figure 4. MIDSS, domain C. Log of the study variable against log of the auxiliary variable, with unity added to both variables

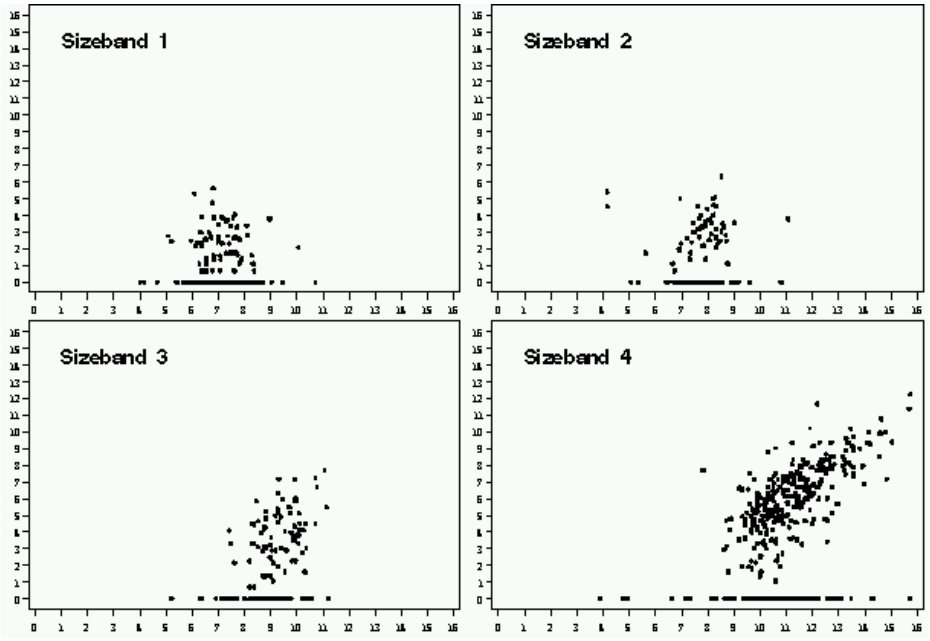


Figure 5. CAPEX, domain U. Log of the study variable against log of the auxiliary variable, with unity added to both variables

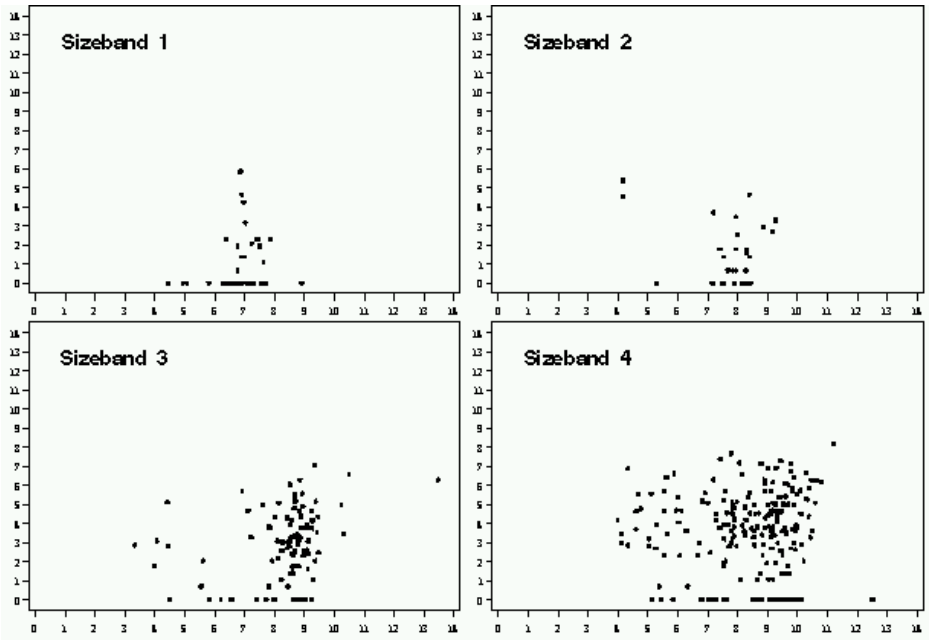


Figure 6. CAPEX, domain V. Log of the study variable against log of the auxiliary variable, with unity added to both variables

In the simulations below we have used the existing strata but allocated the sample on register turnover, with the exception of CAPEX domain V where 'even' sample sizes were chosen. Sample sizes used in simulations are shown in Tables 1 to 5. One thousand samples were drawn from each domain.

Table 1. Sample sizes for the simulated simulated samples, MIDSS domain A

Size-band	N_h	n_h	n_h/N_h %
1	39	9	23
2	33	19	57
3	52	32	62
4	43	43	100
Sum	167	103	62

Table 2. Sample sizes for the samples, MIDSS domain B

Size-band	N_h	n_h	n_h/N_h %
1	73	5	7
2	51	28	54
3	88	67	77
4	74	74	100
Sum	286	174	61

Table 3. Sample sizes for the simulated simulated samples, MIDSS domain C

Size-band	N_h	n_h	n_h/N_h %
1	206	59	29
2	129	13	10
3	305	128	42
4	213	213	100
Sum	853	413	48

Table 4. Sample sizes for the samples, CAPEX domain U

Size-band	N_h	n_h	n_h/N_h %
1	254	25	10
2	107	24	21
3	133	51	38
4	393	393	100
Sum	887	493	56

Table 5. Sample sizes for the simulated samples, CAPEX domain V

Size-band	N_h	n_h	n_h/N_h %
1	40	10	10
2	33	10	30
3	112	30	27
4	202	202	100
Sum	387	252	65

3.1. Properties of an estimator

We are interested in the following measures.

1. ***Coefficient of variance (CV)***. The ratio of the standard deviation of the simulated point estimates to the true total.
2. ***Bias***. The mean of the absolute errors of the simulated estimates divided by the true total.
3. The ***Coverage probability***. The 95% confidence intervals were ± 1.96 times

the square root of the variance estimates (19) – (23).

4. What proportion of the point estimates that are further away from the true total than 0.675 times the standard error of the point estimates. The constant 0.675 is so chosen that if the estimates are normally distributed then 50% will be ***Non-centred.***
5. The maximum of the absolute differences between the 95% and 5% percentile and the true total, divided by the true total. This has the flavour of a minimax criterion with the survey error as loss function. We label this criterion ***Large Error***.

Unfortunately, there is no hard and fast rule which properties to prioritise. The first three are the traditional properties that together with the MSE often are taken as the guiding rule. Despite the strong position of the MSE, there is some arbitrariness in using squared error loss as the one and only loss function (see also Robert, Hwang, and Strawderman, 1993, in particular the discussion that follows the paper). Based on the statistical adage that most sampling distributions are 'normal in the middle', we might expect close to 50% of the estimates to be Non-centred. Property 5 is particularly important in official statistics where the publication of bad estimates may sometimes lead to great losses for society and may also be detrimental to the reputation of the national statistical institute. I would argue that the criterion Large Error is easier to understand and explain than the CV or the MSE.

3.2. Simulation results

Tables 6 to 10 show the CV and other measures for five domains. In the tables, the type of model group is indicated by a number: 1 for 'within strata', 2 for 'ONS model groups' and 3 for 'over all strata'. For example, as seen in Table 6, the estimator most widely used in ONS business survey estimation, here called Rat_2, gives poorer CV than does the expansion estimator, HT, for four out of five domains. Some other observations are listed after each table.

Table 6. Per cent coefficient of variation (CV) for five domains.

	MIDSS			CAPEX	
	A	**B**	**C**	**U**	**V**
HT	2.42	0.92	1.61	1.13	6.62
Rat_1	1.51	1.29	1.05	1.24	15.46
Rat_2	1.74	1.29	1.16	1.15	15.6
Rat_3	1.83	1.34	1.17	1.14	24.83
Reg/1.0_1	1.52	1.28	1.03	1.42	14.01
Reg/1.0_2	1.72	1.28	1.15	1.16	14.2
Reg/1.0_3	1.83	1.34	1.16	1.14	7.94
Reg/1.5_1	1.7	1.38	1.07	1.4	21.74

	MIDSS			CAPEX	
	A	**B**	**C**	**U**	**V**
Reg/1.5_2	1.78	1.36	1.21	1.41	42.11
Reg/1.5_3	1.83	1.41	1.2	1.14	19.35
Local/f40	1.87	1.36	1.19	1.13	6.42
Local/s20	1.83	1.36	1.07	1.14	6.69
RobReg/f40_1	1.87	1.49	1.06	1.19	19.54
RobReg/f40_2	1.83	1.37	1.17	1.27	45.85
RobReg/f40_3	1.82	1.42	1.17	1.13	12.24
Logn/pr_1	1.59	362619	0.96	8E44	..
Logn/pr_2	1.26	1.09	1.02	0.49	6.95
Logn/pr_3	0.77	0.81	0.58	0.33	4.47
Logn/log_1	1.71	379092	0.98	1E45	..
Logn/log_2	1.38	1.1	1.05	0.51	7.01
Logn/log_3	1.03	0.82	0.6	0.38	4.57
RLogn/pr_1	1.67	525230	0.85	8E44	..
RLogn/pr_2	1.45	1.16	0.8	0.58	11.83
RLogn/pr_3	0.86	0.87	0.46	0.26	6.04

1. The Logn and RLogn estimators fitted within strata broke down for several domains. The bias is rather large for other domains as well (Table 7). Consequently, Logn does not perform well in terms of root MSE (not shown here). The reason that Logn and RLogn ‘ within strata’ broke down for three domains is that few units are sampled from sizeband 1 (Tables 2, 4 and 5), many of which may be zero. As all three mixture model estimators only use positive values of the study variable to fit a lognormal model, the fit will be very unstable for small samples. The reason for the poor performance does not seem to be lack of lognormality. The fit is largely rather good for the MIDSS, and those subsets with poor model fit do not correspond to subsets that generate poor estimates.
2. Among design-based estimators it is HT that gives the smallest CV for several domains. The reason is poor correlation between study and auxiliary variables. This lack of correlation arises either through outliers (domain B, Figure 3) or through overall weak association (domains U and V, Figures 5 and 6).
3. For weak-association and outlier-prone domains (such as U and V) larger groups give smaller CV. The opposite is true for the MIDDS domains.
4. In terms of CV, RobReg is among the worst estimators for several domains, including domain V.
5. Local is among the best for the CAPEX, and not far worse than Rat and Reg/1.0 for the MIDSS (between 5% and 24% higher CV than the best Rat or Reg/1.0).
6. In domain V, Figure 6, the extreme-leverage observation in sizeband 3 causes extrapolation far beyond the sample range for all samples without this

observation. The result is unstable estimates for virtually all estimators except the HT estimator, which is per construction insensitive to the auxiliary variable.

7. Reg/1.5 is worse than Reg/1.0 throughout, and far worse for domain V. This is rather surprising considering that the model underlying Reg/1.5 fits data better than that of Reg/1.0.

Table 7. Bias for five domains (per cent of true total)

	MIDSS			CAPEX	
	A	**B**	**C**	**U**	**V**
HT	0.06	-0.05	-0.01	-0.01	0.07
Rat_1	0.42	0.22	0.11	-0.02	9.91
Rat_2	0.12	0.09	0.06	-0.01	9.20
Rat_3	0.04	-0.01	0.01	-0.01	7.41
Reg/1.0_1	0.42	0.27	0.07	-0.22	4.93
Reg/1.0_2	0.13	0.09	0.06	-0.04	-2.80
Reg/1.0_3	0.04	-0.01	0.01	-0.01	0.88
Reg/1.5_1	0.25	0.12	0.05	-0.16	1.62
Reg/1.5_2	0.09	0.01	0.04	-0.10	-2.82
Reg/1.5_3	0.05	-0.01	0.02	-0.02	0.23
Local/f40	0.04	0.01	0.07	-0.01	-1.90
Local/s20	0.06	0.02	0.07	0	-3.04
RobReg/f40_1	0.04	0	0.02	-0.06	0.07
RobReg/f40_2	0.03	-0.02	0.02	-0.05	4.40
RobReg/f40_3	0.03	-0.02	0.01	-0.01	1.04
Logn/pr_1	1.11	14700	-0.04	2E43	3E167
Logn/pr_2	1.27	0.90	1.42	3.39	7.43
Logn/pr_3	1.72	1.19	1.10	3.72	35.2
Logn/log_1	1.02	13300	0.13	4E43	3E167
Logn/log_2	1.19	0.97	1.65	3.46	7.59
Logn/log_3	1.65	1.27	1.32	3.98	35.47
RLogn/pr_1	4.00	19900	-1.07	2E43	3E167
RLogn/pr_2	0.1	0.19	-0.05	3.32	9.75
RLogn/pr_3	0.74	0.51	-0.6	3.32	33.34

1. The bias can be very large for weak-association populations with extreme-leverage points, such as domain V displayed in Figure 6. This is particularly true for Rat applied to a population that calls for a positive intercept. For other populations, linear or not, or outlier prone or not, the bias is negligible for the design-based estimators, including RobReg.

2. Logn/pr and Logn/log tend to give positive bias. This is in accordance with Karlberg's (2000) empirical findings. Rlogn/pr seems rather better in this respect.
3. The bias is often slightly larger for small model groups but still negligible for all domains but one.

Table 8. Coverage probability, in per cent, for five domains.

	MIDSS			CAPEX	
	A	**B**	**C**	**U**	**V**
HT	89.0	92.7	90.6	91.5	87.9
Rat_1	73.8	63.9	90.8	85.6	73.6
Rat_2	80.6	65.1	91.2	89.2	68.8
Rat_3	79.9	64.7	93.2	85.9	31.7
Reg/1.0_1	72.7	63.7	91.1	89.6	84.8
Reg/1.0_2	80.6	65.1	91.1	92.1	86.3
Reg/1.0_3	80.0	64.7	93.3	85.9	90.2
Reg/1.5_1	80.6	65.3	91.2	92.5	90.1
Reg/1.5_2	81.1	65.5	91.9	96.3	85.0
Reg/1.5_3	80.2	64.7	93.1	88.0	87.4
Local/f40	79.4	64.7	93.7	85.3	79.4
Local/s20	78.9	64.7	94.4	85.2	75.5
RobReg/f40_1	79.4	64.6	92.9	84.2	55.2
RobReg/f40_2	80.0	64.7	93.8	81.7	36.9
RobReg/f40_3	80.2	64.7	93.4	85.5	63.7

1. The coverage probability is poor in many cases. No estimator except the HT estimator gives acceptable coverage for all domains. The reason is the non-normality of the estimates for many domains, in particular B. The sample distribution is bimodal for this domain. The main reason for this to happen is the high leverage point in sizeband 3 visible in Figure 3.
2. If the population is linear (such as the MIDSS domains, Figures 2 to 4), then 'within stratum' model groups seem worse than larger model groups, in terms of coverage probability. This makes the lower CV for 'within stratum' a moot point.
3. The variance estimator for RobReg seems unreliable.

Table 9. Per cent Non-centred estimates for five domains.

	MIDSS			CAPEX	
	A	**B**	**C**	**U**	**V**
HT	52	47	51	36	53
Rat_1	55	62	51	34	61
Rat_2	56	64	51	36	67
Rat_3	57	71	50	36	80
Reg/1.0_1	52	58	52	27	48
Reg/1.0_2	56	63	51	35	47
Reg/1.0_3	57	71	50	36	54
Reg/1.5_1	55	67	53	33	55
Reg/1.5_2	56	70	50	35	40
Reg/1.5_3	56	71	51	36	45
Localf40	56	68	50	36	56
Local/s20	58	68	49	37	57
RobReg/f40_1	54	66	51	36	39
RobReg/f40_2	56	72	51	37	27
RobReg/f40_3	56	72	50	36	46
Logn/pr_1	57	0	49	0	0
Logn/pr_2	68	41	76	100	62
Logn/pr_3	94	74	89	100	100
Logn/log_1	56	0	50	0	0
Logn/log_2	64	42	81	100	63
Logn/log_3	86	79	93	100	100
RLogn/pr_1	51	0	78	0	0
RLogn/pr_2	59	57	51	100	46
RLogn/pr_3	66	39	78	100	100

Note: a point estimate is Non-centred if it is further away from the true total than 0.675 times its standard error.

1. The distributions of the estimates are clearly non-normal for domains B, U, and V.
2. Most of the design-based estimators are similar in terms of the Non-centred criterion. However, HT and RobReg stand out, giving equal or better performance.

Table 10. Per cent estimates with Large Error for five domains.

	MIDSS			CAPEX	
	A	**B**	**C**	**U**	**V**
HT	4.1	1.5	2.7	1.2	11.5
Rat_1	2.9	2.3	2	1.2	34.8
Rat_2	2.9	2.2	2.1	1.2	33.7
Rat_3	2.8	2.1	1.9	1.2	33.5
Reg/1.0_1	2.9	2.4	1.9	1.2	29.5
Reg/1.0_2	2.8	2.1	2.1	1.2	17
Reg/1.0_3	2.8	2.1	1.9	1.2	13.7
Reg/1.5_1	3	2.3	1.9	1.6	33.7
Reg/1.5_2	2.8	2.2	2.1	1.7	89.4
Reg/1.5_3	2.8	2.2	2	1.2	36
Local/f40	2.8	2.1	2	1.2	8.8
Local/s20	2.7	2.2	2	1.2	25
RobReg/f40_1	2.8	2.2	1.9	1.2	8.1
RobReg/f40_2	2.8	2.2	1.9	1.2	8.1
RobReg/f40_3	2.8	2.2	1.9	1.2	8.1
Logn/pr_1	3.9	2.7	1.6	3.9	31.9
Logn/pr_2	3.4	2.8	3.2	4.2	19.7
Logn/pr_3	3	2.5	2.1	4.3	42.5
Logn/log_1	3.9	2.7	1.9	4.1	25.7
Logn/log_2	3.4	3	3.4	4.3	20.2
Logn/log_3	3.1	2.6	2.3	4.7	42.5
RLogn/pr_1	3.2	3.4	0.5	4.6	374.7
RLogn/pr_2	2.4	2.1	1.3	4.3	33.1
RLogn/pr_3	2.1	1.9	0.2	3.8	43.8

Note: Large Error is defined as the maximum of the absolute differences between the 95% and 5% percentile and the true total, divided by the true total.

1. In terms of the Large error criterion, RobReg is the best estimator for domain V and no worse than any other estimator for other domains. Local/f40 also performs well.
2. The Large Error and Non-centred criteria combined show that the distribution of 'within stratum' estimates are both more peaked and fat-tailed than the distribution for larger model groups. Again, this makes the lower CV for 'within stratum' a moot point.

4. Discussion

We have conducted a simulation study of estimation in business surveys and contrasted GREGs with a local regression estimator and a robust regression

estimator. We measured each estimator with each of three type of model grouping, where relevant, against five criteria. Three of the criteria are conventional (bias, variance and coverage probability), whereas the other two measured aspects of the absolute error: the proportion of the estimates that were close to the true value and the proportion that where very far from the true value. Some general conclusions are:

1. There is no estimator that is *the best*. It all depends on the use of the estimates and on the population. Different criteria will be more important for different uses. The users, however, should not decide what estimators are being used. The users may change but the national statistical institute cannot afford to flit.
2. The estimators that work reasonably well across all populations are the expansion estimator, Reg/1.0 fitted across all strata and the Local regression estimators. In particular, there seems to be no reason to prefer the ratio estimator to Reg/1.0.
3. The standard way of constructing confidence intervals (1.96 times the standard error, estimated with formulas such as those in Sec 2.4) often gives poor coverage. If the main aim is good confidence intervals then the expansion estimator is preferable, although the price to pay will be wide intervals. There is a need for more research into estimation of confidence intervals.
4. Fitting models within strata (leading to estimators such as the separate ratio or regression estimator) tends to give small CVs, but fitting models across strata tends to make estimates more robust.
5. High leverage points need to be addressed. Hedlin et al. (2001) suggest post-stratification where the HT estimator is applied to the 20 or so units with the largest values of the auxiliary and that e.g. some GREG estimator that does use auxiliary information is applied to post-strata without high leverage points.

Other conclusions that concern specific estimators are:

- The choice of bandwidth for local regression estimators does not seem overly sensitive. This is rather surprising considering that choice of bandwidth is regarded as sensitive in applications of local regression outside the sample survey area.
- The robust regression estimator and one of the local regression estimators are superior if the aim is to minimise the proportion of estimates that are very far from the true value in absolute terms. This is particularly important in official statistics.
- The model-based mixture model estimator is bias prone and will give poor estimates for some populations. Robust estimation of the variance parameter seems to be an approach that reduces some of the problems. Also, if the model is fitted across strata, including the completely enumerated stratum, the parameter estimation tends to be more reliable. However, like the robust regression estimators, this estimator is not linear, nor has it the internal

consistency property.

- The regression estimator that is associated with the best model (with variance about the regression line proportional to the auxiliary raised to 1.5) is more erratic than the regression estimator modelled on a variance proportional to the auxiliary. The reason was seen to be variance in the bias adjustment term, i.e., the second term in (1) and (4). This term is non-zero only for the former estimator.

Acknowledgments

The author is grateful to several members of staff at the U.K. Office for National Statistics for valuable comments and for providing data.

REFERENCES

ALBERT, A. and ANDERSON, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. Biometrika, 71, 1-10.

BELSLEY, D.A., KUH, E., and WELSCH, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

BREIDT, F.J. and OPSOMER, J.D. (2000). Local Polynomial Regression Estimation in Survey Sampling. The Annals of Statistics, 28, 1026-1053.

CHAMBERS, R.L., DORFMAN, A.H., and WEHRLY, T.E. (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. Journal of the American Statistical Association, 88, 268-277.

CHAMBERS, R.L. and DUNSTAN, R. (1986). Estimating Distribution Functions from Survey Data. Biometrika, 73, 597-604.

CHAMBERS, R.L. and KOKIC, P.N. (1993). Outlier Robust Sample Survey Inference. Bulletin of the International Statistical Institute, Invited Papers, 69-86.

COCHRAN, W.G. (1977). Sampling Techniques, 3rd ed. New York: Wiley.

HEDLIN, D., FALVEY, H., CHAMBERS, R., and KOKIC, P. (2001). Does the Model Matter for GREG Estimation? A Business Survey Example. Journal of Official Statistics, 17, 527-544.

KARLBERG, F. (2000). Survey Estimation for Highly Skewed Populations in the Presence of Zeroes. Journal of Official Statistics, 16, 229-241.

KOKIC, P.N. and BELL, P.A. (1994). Optimal Winsorizing Cutoffs for a

Stratified Finite Population Estimator. Journal of Official Statistics, 10, 419-435.

KUK, A.Y.C. and WELSH, A.H. (2001). Robust Estimation for Finite Populations Based on a Working Model. Journal of the Royal Statistical Society, B, 63, 277-292.

ROBERT, C.P., HWANG, J.T.G., and STRAWDERMAN, W.E. (1993). Is Pitman Closeness a Reasonable Criterion? (with Discussion). Journal of the American Statistical Association, 88, 57-76.

RYAN, T.P. (1997). *Modern Regression Methods. New York: Wiley.*

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

SILVA, P.L.D.N. and SKINNER C.J. (1997). Variable Selection for Regression Estimation in Finite Populations. Survey Methodology, 23, 23-32.

SUGDEN, R.A. and SMITH, T.M.F. (2002). Exact Linear Unbiased Estimation in Survey Sampling. Journal of Statistical Planning and Inference, 102, 25-38.

VALLIANT, R., DORFMAN, A.H., and ROYALL, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley.

WELSCH, R.E. (1980). Regression Sensitivity Analysis and Bounded-Influence Estimation. In Evaluation of Econometric Models, eds. J. Kmenta and J.B. Ramsey. New York: Academic Press, 153-167.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 969–994

A MULTIPARAMETER PERSPECTIVE ON THE CHOICE OF SAMPLING DESIGN IN SURVEYS

Anders Holmberg[1]

ABSTRACT

At the design and estimation stages of a survey, large survey organizations often use auxiliary information. Technological advances in data capture and a better accessibility of registers open up for an increased and more efficient use of such information. This paper addresses issues of how to use auxiliary information efficiently in sampling from finite populations. Previous results regarding the choice of optimal design are extended to the case of several study variables. We suggest approaches to achieve a high overall efficiency, and compare these approaches with single-variable routines, often used by practising survey statisticians.

***Key words* and Phrases:** Auxiliary information, GREG Estimator, Optimal designs, Survey planning.

1. Introduction

A sample survey is in general taken with the purpose of estimating a large set of parameters $\theta_1, \theta_2, \ldots, \theta_i, \ldots$ (totals, means, ratios, medians, Gini coefficients etc.) of a finite population. The most important of these parameters form the basis for planning the survey, and the survey statistician's task is to find an efficient combination of sampling design and estimator vector, i.e. such that the final choice results in 'small' mean square error for each estimator $\hat{\theta}_i$.

When there is only one parameter to estimate, say a population total, the statistician might start by choosing a suitable member, $\hat{\theta}$, of a specific class of estimators known to have good properties, e.g. GREG (generalized regression) estimators, followed by an attempt to find a sampling design that minimizes the mean square error of $\hat{\theta}$ or variance if $\hat{\theta}$ is unbiased.

[1] Statistics Sweden, Department of Research and Development, SE-701 89 Örebro, Sweden.
e-mail: Anders.Holmberg@scb.se

There is no simple straightforward generalization of this single parameter approach to the multiparameter case. However, in this paper we will discuss a couple of approaches that might be useful to achieve high overall efficiency.

The paper has the following structure. In the following preliminaries, we give the background to the problem, introduce our basic notations and specify the survey situation that is considered. In section 2, we treat the single parameter case and give an overview of results on the choice of 'optimal' designs. It lays the basis of section 3, which contains results of our approaches concerning the multiparameter case. Both section 2 and section 3 naturally leads to treatment of without replacement probability proportional-to-size sampling, (πps sampling), combined with GREG estimation. Since the survey statistician needs feasible methods to implement theory, and since relatively recent progress in the area of πps-sampling has been made, a short overview is provided in section 4. Comparisons between our suggested approaches are made in section 5. In section 6, a further extension of the multiparameter approaches is outlined, and, finally, our conclusions and recommendations are given in section 7.

1.1. Background and notation

1.1.1. General definitions

Let $U = \{1,\ldots,k,\ldots,N\}$ be a finite population consisting of N elements. We consider Q unknown study variables $y_1,\ldots,y_q,\ldots,y_Q$. Let y_{qk} denote the value of y_q for population element k. To select a set sample $s \subseteq U$ of size n_s, a without replacement sampling design, $p(\cdot)$, will be used. We denote the first-order inclusion probabilities by $\pi_k \ (k = 1,\ldots,N)$ and the second-order inclusion probabilities by $\pi_{kl} \ (k,l = 1,\ldots,N)$.

The vector of the most important parameters, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_i,\ldots,\theta_I)'$, often consists of functions of the population totals, i.e. $\boldsymbol{\theta} = (\, f_1(\mathbf{t}),\, f_2(\mathbf{t}),\ldots,\, f_i(\mathbf{t}),\ldots,\, f_I(\mathbf{t}))'$ where $\mathbf{t} = (t_{y_1},\ldots,t_{y_q},\ldots,t_{y_Q})'$ and $t_{y_q} = \sum_{k\in U} y_{qk} = \sum_U y_{qk}$. Henceforth, we will consider the case where $\boldsymbol{\theta} = \mathbf{t}$ (i.e. $I = Q$.)

A strategy, $\Omega_{p,\hat{\theta}}$, is defined as a combination of a sampling design and an estimator vector $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1,\ldots,\hat{\theta}_i,\ldots,\hat{\theta}_I\right)'$, i.e. in our case where $\hat{\boldsymbol{\theta}} = \hat{\mathbf{t}}$, we have $\Omega_{p,\hat{t}} = [p(*),\hat{t})]$.

Many results in survey theory address the problem of finding an efficient strategy. Those results rely on the availability of auxiliary information (e.g. the literature on optimum allocation in stratified sampling, on πps sampling designs or on estimators using auxiliary variables.) This paper assumes that there are P auxiliary variables accessible at the planning stage. They are denoted $u_1,\ldots,u_p,\ldots,u_P$ and their values u_{pk} $\left(p = 1,\ldots,P\right)$ are known for every element *k* in the population.

1.1.2 Model assisted estimation (the GREG estimator)

When we plan our survey strategy, we try to use the auxiliary information in the choice of *design* as well as in the choice of *estimator*, in such a way that the sampling error of $\hat{\mathbf{t}}$ becomes as small as possible. Statistical models can be used as an aid in this planning process. Hence, we assume that the statistician has useful a priori knowledge about the relations between the study variables and the auxiliary variables.

We presume the structure of these relations makes it relevant to formulate linear models, ξ_q, $\left(y_{qk} = \mathbf{x}'_{qk}\boldsymbol{\beta}_q + \varepsilon_{qk}\right)$ for the study variables, with $E_{\xi_q}\left(\varepsilon_{qk}\right) = 0$, $V_{\xi_q}\left(\varepsilon_{qk}\right) = \sigma^2_{qk}$ and $E_{\xi_q}\left(\varepsilon_{qk}\varepsilon_{ql}\right) = 0$ $\left(k \neq l\right)$, i.e.,

$$E_{\xi_q}\left(y_{qk}\right) = \mathbf{x}'_{qk}\boldsymbol{\beta}_q \quad \left(k = 1,\ldots,N\right)$$

$$V_{\xi_q}\left(y_{qk}\right) = \sigma^2_{qk} \quad \left(k = 1,\ldots,N\right) \tag{1}$$

where $\mathbf{x}'_{qk} = \left(x_{1_q k},\ldots,x_{j_q k},\ldots,x_{J_q k}\right)$ is a suitable set of J_q (positive) auxiliary variables formed from $u_1,\ldots,u_p,\ldots,u_P$, and $\boldsymbol{\beta}_q = \left(\beta_{1_q},\ldots,\beta_{j_q},\ldots,\beta_{J_q}\right)'$ are model parameters. The values $\sigma^2_{q1},\ldots,\sigma^2_{qN}$ are considered known, although knowledge up to a constant multiplier sometimes is sufficient.

The population total of y_q, $t_{y_q} = \sum_U y_{qk}$ can be estimated by the GREG estimator, which is defined as

$$\hat{t}_{y_q r} = \hat{t}_{y_q \pi} + \left(\mathbf{t}_{x_q} - \hat{\mathbf{t}}_{x_q \pi}\right)' \hat{\mathbf{B}}_q \tag{2}$$

Here, $\hat{t}_{y_q\pi} = \sum_{k\in s} y_{qk}/\pi_k = \sum_s y_{qk}/\pi_k$ is the well known Horvitz-Thompson or π estimator, $\mathbf{t}_{x_q} = \left(t_{x_{1q}},\ldots,t_{x_{jq}},\ldots,t_{x_{J_q}}\right)'$ is a J_q-dimensional vector of x_q totals, $\hat{\mathbf{t}}_{x_q\pi}$ is a vector of the corresponding π estimators and

$$\hat{\mathbf{B}}_q = \left(\sum_s \frac{\mathbf{x}_{qk}\mathbf{x}'_{qk}}{c_{qk}\pi_k}\right)^{-1} \sum_s \frac{\mathbf{x}_{qk}y_{qk}}{c_{qk}\pi_k} \tag{3}$$

is an estimated vector of regression coefficients, where c_{qk} is a suitable constant. Moreover, the Taylor expansion variance of $\hat{t}_{y_q r}$ is given by

$$V_T(\hat{t}_{y_q r}) = \sum_{k\in U}\sum_{l\in U} \frac{\left(\pi_{kl} - \pi_k\pi_l\right)}{\pi_k\pi_l} E_{qk}E_{ql} \tag{4}$$

where $E_{qk} = y_{qk} - \mathbf{x}'_{qk}\mathbf{B}_q \quad (k = 1,\ldots,N)$ are population fit residuals, with $\mathbf{B}_q = \left(\sum_U \mathbf{x}_{qk}\mathbf{x}'_{qk}/c_{qk}\right)^{-1}\sum_U y_{qk}\mathbf{x}'_{qk}/c_{qk}$ a finite population regression coefficient. (Details of GREG estimation are given in Särndal, Swensson and Wretman (1992) sections 6.4-6.7.)

Henceforth, the results in this paper concern the family of GREG estimators which includes several well known estimators used in practice, (e.g. the post-stratified estimator and the common ratio estimator.) A GREG estimator is approximately unbiased even with poorly fitted models, but with a strong relationship between y and $\mathbf{x}$, and a fair knowledge of that relationship, the GREG estimator will outperform the π estimator as far as efficiency is concerned. However, other types of estimators could also be mentioned when we discuss strategy selection using auxiliary information.

1.1.3. Alternatives to the GREG estimator

An estimator related to the GREG estimator is the 'optimal' regression estimator', $\hat{t}_{y_q or}$, (Rao (1992, 1994, 1997), Cassady and Valiant (1993) and Montanari (1987, 1998)),

$$\hat{t}_{y_q or} = \hat{t}_{y_q\pi} + \left(\mathbf{t}_{x_q} - \hat{\mathbf{t}}_{x_q\pi}\right)'\tilde{\mathbf{B}}_q$$

where $\tilde{\mathbf{B}}_q = \left[\hat{V}(\hat{\mathbf{t}}_{x_q\pi})\right]^{-1}\hat{C}(\hat{\mathbf{t}}_{x_q\pi},\hat{t}_{y_q\pi})$, with $\hat{V}(\hat{\mathbf{t}}_{x_q\pi})$ and $\hat{C}(\hat{\mathbf{t}}_{x_q\pi},\hat{t}_{y_q\pi})$ being unbiased estimators for $V(\hat{\mathbf{t}}_{x_q\pi})$ and $C(\hat{\mathbf{t}}_{x_q\pi},\hat{t}_{y_q\pi})$ of dimensions $J_q \times J_q$ and

$J_q \times 1$ respectively. (Expressions for the well known Horvitz-Thompson or alternatively the Sen-Yates-Grundy variants of $\hat{V}(\hat{\mathbf{t}}_{x_q\pi})$ and $\hat{C}(\hat{\mathbf{t}}_{x_q\pi}, \hat{t}_{y_q\pi})$ can be found in Särndal et al. pp 44-45, 170.) To use the variance and covariance estimators in point estimation can be very impractical and sometimes lead to trouble. Nevertheless, Montanari (1998) discusses some situations when $\hat{t}_{y_q or}$ can be preferred over the GREG estimator.

Another more useful and wider family of estimators, are the calibration estimators described in Deville and Särndal (1992), Lundström and Särndal (1999) and Estevao and Särndal (2000). They are asymptotically equivalent to the GREG estimator, and they are appealing to practitioners in attempts to reduce non-response bias.

Nonetheless, in the following neither calibration estimators nor $\hat{t}_{y_q or}$, will be discussed. This restricts our strategy concept to strategies where the estimator is a member of the GREG estimator family, (different GREG estimators for different parameters are allowed.) Since the above estimators are related to the GREG estimator, we believe that this is a mild restriction. In the planning stage of a multiparameter survey, the choice of design, which affects all parameter estimates, is likely to be more important than the choice between related estimators. Therefore, if we can find an 'optimal' *design* for an efficient estimator such as a GREG estimator, such a design is apt to work well combined with the other estimators as well. (As to the choice of a *specific* GREG estimator, it is henceforth a tacit understanding that when a model like (1) is explicitly presented, the model is satisfactory, and the information given is good enough for the statistician to make a suitable GREG estimator choice.)

Note that we consider design-based inference, but we try to make as efficient use of supplementary information as possible through models. Our approach will be model assisted, and we use models to assist in our design selection and in choosing estimators for $\mathbf{t}$. In the next section we reproduce results on finding an 'optimal' design for a GREG estimator in the single parameter case.

2. Selecting an optimal design in the single parameter case

If we consider GREG estimation, the question of which strategy to choose can be limited to the issue of finding a design that minimizes $V(\hat{t}_{y_q r})$. However, direct minimization of $V(\hat{t}_{y_q r})$ is impossible, but given a model ξ_q, and utilizing the statistical properties of the model errors ε_{qk}, we can try to minimize the anticipated variance, (i.e. the variance over both the model ξ_q and the design *p*, e.g. see Isaki and Fuller (1982)).

$$E_{\xi_q} E_p \left[\left(\hat{t}_{y_q r} - t_{y_q} \right)^2 \right] - \left[E_{\xi_q} E_p \left(\hat{t}_{y_q r} - t_{y_q} \right) \right]^2$$

If the model ξ_q is well specified, then an approximation to the anticipated variance can be written as (see Särndal et al. pp 450-451),

$$ANV_q(\hat{t}_{y_q r}) = \sum_U \left(\pi_k^{-1} - 1 \right) \sigma_{qk}^2 \tag{5}$$

For a sampling design $p(*)$ such that $E_p(n_s) = n$, Result 12.1.1. in Särndal et al. show that a near 'optimal' design, i.e. a design that minimizes (5), is such that the first-order inclusion probabilities for $k = 1, \ldots, N$ are given by

$$\pi_k = \pi_{q(opt)k} = n\sigma_{qk} / \sum_U \sigma_{qk} \tag{6}$$

and the minimum of $ANV_q(\hat{t}_{y_q r})$ is

$$\begin{aligned} ANV_{q\min}(\hat{t}_{y_q r}) &= \sum_U \left(\pi_{q(opt)k}^{-1} - 1 \right) \sigma_{qk}^2 \\ &= \frac{1}{n} \left(\sum_U \sigma_{qk} \right)^2 - \sum_U \sigma_{qk}^2 \end{aligned} \tag{7}$$

Hence, we obtain an 'optimal' design by choosing $\pi_k \propto \sigma_{qk}$, i.e. the statistician should use a πps design with σ_{qk} as size measures. For example, if $\sigma_{qk}^2 = \sigma_q^2 u_{qk}^{\gamma_q}$, where σ_q^2 is a constant (possibly unknown) and $u_q \in \{u_1, \ldots, u_p, \ldots, u_P\}$ is one of the auxiliary variables, this result suggests that by using a well chosen GREG estimator, and choosing a design where $\pi_k \propto u_{qk}^{\gamma_q / 2}$, we obtain a near 'optimal' strategy for estimating t_{y_q}.

We will return to the issues of implementing a πps design in section 4, but from here on, designs where π_k are proportional to some known size measure, *z*, are referred to as $\pi ps(z)$ designs. For the moment, we conclude that from a theoretical view, the combination of GREG estimation and a $\pi ps(\sigma)$ design is near 'optimal' with respect to minimizing an approximation to the anticipated variance. In the single parameter case, similar conclusions can be made, (for fixed size designs), from the results in Cassel, Särndal and Wretman (1976, 1977), (Theorem 1 or Theorem 4.1 respectively) and Theorem 2.1 in Rosén (2000a).

The results above can be helpful for survey planning but the limitations are obvious. As for many other optimality results, focus is on a single study variable, which is insufficient for a practising survey statistician having to deal with several

parameters of importance. (For thorough reviews on optimization problems in choosing sampling designs for surveys, see Rao (1979) and Bellhouse (1984).)

3. Selecting a 'best' overall design in the multi-parameter case

More realistically, suppose that we want to estimate $\mathbf{t} = \left(t_{y_1}, \ldots, t_{y_q}, \ldots, t_{y_Q}\right)'$, where $Q \geq 2$, and let the relative importance of the parameters be reflected by a set of importance weights H_q, $\left(\sum_{q=1}^{Q} H_q = 1\right)$. Moreover, suppose that a good choice of GREG estimator can be made for each population total t_{y_q}. Then, the statistician's task is to find a design that in some sense could be considered optimal for all parameters.

However, one problem now is that there is no - in contrast to the single parameter case – single well-defined meaning of 'optimality'. A design that is optimal or close to optimal in the single parameter case might not be the best choice in an overall multiparameter sense. For example, suppose a diligent statistician with a specified amount of auxiliary information, time and skill, should use the 'optimal' design result above to seek what he or she a priori believes to be the theoretically 'optimal' design for every parameter, t_{y_q}, $\left(q = 1, \ldots, Q\right)$, separately. Most likely, he then would find design solutions that differ, and only one of them can be chosen. The statistician in such a situation is forced to seek a compromise design, which he believes works reasonably well for all parameters to be estimated.

Here, three different compromise approaches that can be used to plan the selection of a design in the multiparameter case are discussed. They all have different minimization criteria, and all are in a sense extensions of the result from the previous section. The first two approaches (A and B) are appealing at a first glance, but they have some built in scaling problems that are avoided in approach C. Since the proofs are carried out in a similar way, we will only provide details for the third approach (approach C).

3.1. Approach A: Minimizing a weighted sum of variances

In the multiparameter situation, a straightforward criterion for selecting the best overall design for estimating $\mathbf{t}$ is to minimize the trace of $V(\hat{\mathbf{t}}_r)$, i.e. minimizing the sum $\sum_{q=1}^{Q} V(\hat{t}_{y_q r})$, or if we like to attach importance weights H_q, minimizing $\sum_{q=1}^{Q} H_q V(\hat{t}_{y_q r})$. Since $\sum_{q=1}^{Q} V(\hat{t}_{y_q r})$ $\left(q = 1, \ldots, Q\right)$ is unknown, we could instead look for the design that, under the restriction

$\sum_U \pi_k = n$ and assumed models ξ_q, $(q = 1,\ldots,Q)$ (see (1)), minimizes a weighted sum of approximated anticipated variances, i.e. a design that minimizes $SANV(\hat{\mathbf{t}}_r) = \sum_{q=1}^{Q} H_q ANV_q(\hat{t}_{y_q r})$. Rewriting this weighted sum as $SANV(\hat{\mathbf{t}}_r) = \sum_U \left(\pi_k^{-1} - 1\right) \sum_{q=1}^{Q} H_q \sigma_{qk}^2$, and using the Cauchy-Schwarz inequality directly yields Result 3.1.

Result 3.1. *A sampling design* $p(*)$ *with the expected sample size* $E_p(n_s) = n$ *that minimizes* $SANV(\hat{\mathbf{t}}_r)$*, is such that the first order inclusion probabilities for* $k = 1,\ldots,N$ *are determined by*

$$\pi_k = \pi_{(A)k} = \frac{n\sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2}}{\sum_U \sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2}} \tag{8}$$

Clearly, a design with $\pi_k = \pi_{(A)k}$ is a compromise that considers all the involved parameters and their importance. However, for $q = 1,\ldots,Q$, $ANV_{qA}(\hat{t}_{y_q r})$ will differ from $ANV_{q\min}(\hat{t}_{y_q r})$, since, in general, the ratios in the inequality below will be larger than 1.

$$\frac{ANV_{qA}(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} = \frac{\sum_U \left(\pi_{(A)k}^{-1} - 1\right)\sigma_{qk}^2}{\sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2} \geq 1 \qquad q = 1,\ldots,Q$$

The sizes of the Q ratios between $ANV_{qA}(\hat{t}_{y_q r})$ and $ANV_{q\min}(\hat{t}_{y_q r})$ will depend on H_q, and σ_{qk}^2. In addition, since $V_{\xi_q}(y_{qk}) = \sigma_{qk}^2$ for $q = 1,\ldots,Q$, the resulting design will have properties that are dependent on the measurement scales of the variables involved in the Q models.

3.2. Approach B: Minimizing a weighted sum of relative variances

Another measure often used by statisticians in survey planning is the coefficient of variation of the estimators. In approach B we use a relative variance, $RV_q(\hat{t}_{y_q r}) = V_q(\hat{t}_{y_q r}) / t_q^2$. Suppose that we want to minimize $\sum_{q=1}^{Q} H_q RV_q(\hat{t}_{y_q r})$. Again, $V_q(\hat{t}_{y_q r})$ is unattainable but we can use $ANV_q(\hat{t}_{y_q r})$. Our approximated relative variance measure then becomes

$ANRV_q(t_{y_q r}) = ANV_q(\hat{t}_{y_q r})/t_q^2$, and we seek the design that minimizes $SANRV(\hat{\mathbf{t}}_r) = \sum H_q ANV_q(\hat{t}_{y_q r})/t_q^2$.

Result 3.2. *A sampling design* $p(*)$ *with the expected sample size* $E_p(n_s) = n$ *that minimizes* $SANRV(\hat{\mathbf{t}}_r)$, *is such that the first order inclusion probabilities for* $k = 1,\ldots,N$ *are determined by*

$$\pi_k = \pi_{(B)k} = \frac{n\sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2 / t_q^2}}{\sum_U \sqrt{\sum_{q=1}^{Q} H_q \sigma_{qk}^2 / t_q^2}} \tag{9}$$

Since we are minimizing a sum of relative measures, this approach is less sensitive to different scales of the involved variables than approach A. However, if $\sigma_{qk}^2 = \sigma_q^2 f_q(u_{qk})$, the constant multiplier σ_q^2 does not cancel out and it may differ in size between the terms in $\sum_{q=1}^{Q} H_q \sigma_{qk}^2 / t_q^2$. Moreover, the requirements on the auxiliary information with this approach are extremely high. In addition to σ_{qk}^2 the approach involves knowledge of the parameters, t_q $(q = 1,\ldots,Q)$, that we want to estimate! In practice, t_q^2 (as well as σ_{qk}^2) must be substituted by planning values ('guesstimates') $\dot{t}_q^2$, and poor guesses of t_q^2 can have large effects on the design properties.

Instead, we propose a less scale dependent approach, which also relates to how much we lose in precision compared to the single parameter 'optimal' designs, given in section 2.

3.3. Approach C: Minimizing a weighted sum of relative efficiency losses

Variance ratios are often used to compare the efficiency of strategies. This principle is used in approach C. As a background to motivate our measure and minimization criterion, let $V(\hat{t})_{\Omega_{opt}}$ denote the estimator variance of an optimal strategy (i.e. the strategy gives the smallest possible sampling error when estimating *t*) and let $V(\hat{t})_{\Omega_{p,\hat{t}}}$ be the estimator variance of $\hat{t}$ for another strategy $\Omega_{p,\hat{t}}$. The relative loss in efficiency (i.e. variance increase) for one strategy compared to the optimal strategy can then be defined as $REL = \left(V(\hat{t})_{\Omega_{p,\hat{t}}} - V(\hat{t})_{\Omega_{opt}}\right)/V(\hat{t})_{\Omega_{opt}}$. Then, with Q *y*-totals to estimate, the overall (total) relative loss in efficiency is

$$OREL = \sum_{q=1}^{Q} \frac{V(\hat{t}_{y_q})_{\Omega_{p,\hat{t}yq}} - V(\hat{t}_{y_q})_{\Omega_{qopt}}}{V(\hat{t}_{y_q})_{\Omega_{qopt}}} \tag{10}$$

With $\hat{\mathbf{t}} = \hat{\mathbf{t}}_r$ we realize from section 2, that, by using a model ξ_q (5) and (6), we can theoretically derive an 'optimal' design, with $\pi_{q(opt)k}$, $(k = 1,\ldots,N)$, for every q, $(q = 1,\ldots,Q)$. We can also calculate $ANV_{q\min}(\hat{t}_{y_q r})$ (see equation (7)) for every $\hat{t}_{y_q r}$, $(q = 1,\ldots,Q)$.

By letting $ANV_{q\min}(\hat{t}_{y_q r})$ take the place of $V(\hat{t})_{\Omega_{opt}}$ in (10) we can formulate approach C as finding the design that minimizes an approximation to (a weighted) *Anticipated Overall Relative Efficiency Loss*, $(ANOREL)$, here defined for GREG estimators as

$$ANOREL = \sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_q r}) - ANV_{q\min}(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} \tag{11}$$

Result 3.3. *Let* $p(*)$ *be a sampling design with the expected sample size* $E_p(n_s) = n$. *Suppose* $\mathbf{t} = \left(t_{y_1},\ldots,t_{y_q},\ldots,t_{y_Q}\right)'$ *is estimated by* $\hat{\mathbf{t}} = \left(\hat{t}_{y_1 r},\ldots,\hat{t}_{y_q r},\ldots,\hat{t}_{y_Q r}\right)'$ *as defined by (2). The design for which the anticipated overall relative efficiency loss (11) is minimized, is such that the first order inclusion probabilities for* $k = 1,\ldots,N$ *are determined by*

$$\pi_k = \pi_{(C)k} = \frac{n\sqrt{\sum_{q=1}^{Q} H_q \frac{\sigma_{qk}^2}{\sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2}}}{\sum_U \sqrt{\sum_{q=1}^{Q} H_q \frac{\sigma_{qk}^2}{\sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2}}} \tag{12}$$

where $\pi_{q(opt)k}$ *is given by (6). The minimum value of* $ANOREL$ *is then*

$$
\begin{aligned}
ANOREL_{\min} &= \sum_{q=1}^{Q} H_q \frac{ANV_{qC}(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} - \sum_{q=1}^{Q} H_q \\
&= n\sum_{q=1}^{Q} H_q \frac{\sum_U \pi_{(C)k}^{-1}\sigma_{qk}^2 - \sum_U \sigma_{qk}^2}{\left(\sum_U \sigma_{qk}\right)^2 - \sum_U \sigma_{qk}^2} - 1
\end{aligned} \tag{13}
$$

where $ANV_{qC}(\hat{t}_{y_q r}) = \sum_U \left(\pi_{(C)k}^{-1} - 1\right)\sigma_{qk}^2$ *and* $ANV_{q\min}(\hat{t}_{y_q r})$ *is given by (7).*

Proof. Let π_k (where $\pi_k \le 1$ for $k = 1,\ldots,N$) denote the first order inclusion probabilities of a design $p(*)$ where $E_p(n_s) = \sum_U \pi_k = n$. Minimizing (11) is equivalent to minimizing

$$
\sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} =
$$

$$
\sum_{q=1}^{Q} H_q \sum_U \left(\pi_k^{-1} - 1\right) \frac{\sigma_{qk}^2}{\sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2} \tag{14}
$$

Simplifying by letting $\omega_{qk} = \sigma_{qk}^2 \Big/ \sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2$ and changing the order of summation we get,

$$
\sum_{q=1}^{Q} H_q \frac{ANV_q(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} = \sum_U \left(\pi_k^{-1} - 1\right)\sum_{q=1}^{Q} H_q \omega_{qk} \tag{15}
$$

Evaluating the right side of (15) and applying the Cauchy-Schwarz inequality give

$$
\left(\sum_U \pi_k\right)\sum_U \frac{\sum_{q=1}^{Q} H_q \omega_{qk}}{\pi_k} \ge \left(\sum_U \left(\sum_{q=1}^{Q} H_q \omega_{qk}\right)^{1/2}\right)^2
$$

where equality holds if and only if $\pi_k = \pi_{(C)k} \propto \sqrt{\sum_{q=1}^{Q} H_q \omega_{qk}}$. Equation (13) is obtained by inserting $\pi_{(C)k}$ for π_k in the numerator and $\pi_{q(opt)k}$ for π_k in the denominator of (11) and evaluating (using (7) and $\sum_{q=1}^{Q} H_q = 1$.) ;

Remark 1 *Proofs of results 3.1 and 3.2 are derived in a similar manner.*

Approach C has an advantage over A and B. Since the measure that is minimized is based on ratios, the scaling effects of the involved variables are neutralized. This can be observed in example 2 below, where the constant factors σ_q^2 cancel out.

3.3.1. Examples

A simple example illustrates how result 3.3 can be used in practice. For simplicity, we now consider the case when all the parameters have equal importance weights, i.e. $H_q = 1/Q$ for $(q = 1,\ldots,Q)$.

Example 2 In the planning stages, Result 3.3 can be used to select a design which gives us a strategy $\Omega_{p,\hat{t}}$ that will give a low overall relative efficiency loss. Again, suppose we want to estimate $\mathbf{t}$ *,and that we have auxiliary information to formulate models ξ_q, $(q = 1,\ldots,Q)$ that are good descriptions of the (y_q, u_q) scatter-plots.*

$$E_{\xi_q}(y_{qk}) = \beta_{0q} + \beta_{1q}u_{qk} \quad (k = 1,\ldots,N)$$

$$V_{\xi_q}(y_{qk}) = \sigma_{qk}^2 = \sigma_q^2 u_{qk}^{\gamma_q} \quad (k = 1,\ldots,N)$$

Furthermore, with 'guesstimates', $\tilde{\gamma}_q$ of γ_q perhaps taken from previous surveys or subject knowledge, we can for $q = 1,\ldots,Q$ and $k = 1,\ldots,N$ easily calculate $\dot{\pi}_{q(opt)k} = nu_{qk}^{\tilde{\gamma}_q/2} / \sum_U u_{qk}^{\tilde{\gamma}_q/2}$. (Henceforth, we use a dot i.e. $\dot{\pi}$ to emphasize inclusion probabilities that depend on assumed or approximated numerical values of σ_{qk}^2.) Result 3.3 then implies that by using GREG estimators $\hat{\mathbf{t}}_r = \left(\hat{t}_{y_1 r},\ldots,\hat{t}_{y_q r},\ldots,\hat{t}_{y_Q r}\right)'$ as defined in (2) and choosing a design such that the first-order inclusion probabilities are determined by,

$$\pi_k = \dot{\pi}_{(C)k} = \frac{n\sqrt{\sum_{q=1}^{Q} \dfrac{u_{qk}^{\tilde{\gamma}}}{\sum_U \left(\dot{\pi}_{q(opt)k}^{-1} - 1\right) u_{qk}^{\tilde{\gamma}}}}}{\sum_U \sqrt{\sum_{q=1}^{Q} \dfrac{u_{qk}^{\tilde{\gamma}}}{\sum_U \left(\dot{\pi}_{q(opt)k}^{-1} - 1\right) u_{qk}^{\tilde{\gamma}}}}}$$

we expect to get a 'small' overall relative efficiency loss.

The reasoning in this section can also be applied for studying the effect of different sample sizes. For example, calculating $ANV_{q\min}(\hat{t}_{y_q r})$ for $q = 1,\ldots,Q$

with a given *n*, gives us the opportunity to study the sample size, n^*, that is needed for (13) to meet certain constraint criteria.

Example 3 Suppose all parameters are equally important (i.e. $H_q = 1/Q$ *for* $q = 1,\ldots,Q$*), and let* n^* *be the sample size needed so that, on average,* $ANV_q(\hat{t}_{y_q r})$ *does not exceed* $ANV_{q\min}(\hat{t}_{y_q r})$ *by more than* $100\times c\%$. *Furthermore, suppose we begin with a starting value of n and calculate* $ANV_{q\min}(\hat{t}_{y_q r})$ *for every* $q = 1,\ldots,Q$. *If the constraint c is not too strict, we then can calculate the smallest value of* n^* *that satisfies the inequality,*

$$\frac{1}{Q}\sum_{q=1}^{Q}\frac{ANV_q(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} - 1 \quad \leq \quad c$$

$$\text{i.e.} \qquad \sum_{q=1}^{Q}\frac{ANV_q(\hat{t}_{y_q r})}{ANV_{q\min}(\hat{t}_{y_q r})} \quad \leq \quad (1+c)Q \qquad (16)$$

By writing $a_{qk} = \sum_{q=1}^{Q} w_{qk} = \sum_{q=1}^{Q} \sigma_{qk}^2 \Big/ \sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2$ *and* $\pi_{(C)k} = n^* \sqrt{a_{qk}} \,/ \sum_U \sqrt{a_{qk}}$ $\left(k = 1,\ldots,N\right)$*, and using (15), the left hand side of (16) can be written* $\left(\sum_U \sqrt{a_{qk}}\right)^2 / n^* - \sum_U a_{qk}$ *, and after some algebra we get,*

$$n^* \geq \frac{\left(\sum_U \sqrt{a_{qk}}\right)^2}{(1+c)Q + \sum_U a_{qk}}$$

Hence, it is possible to determine the sample size that is needed to meet specifications made on $ANOREL_{\min}$ *in formula (13). (When further elaborating Approach C (see section 6) such calculations might be helpful.)*

Equations (6), (8), (9) and (12) all suggest that the inclusion probabilities should be chosen proportional to some function of $V_{\xi_q}(y_{qk}) = \sigma_{qk}^2$. To apply this we need a good sampling scheme to implement the suggested designs, and we need to have a good idea of the values of σ_{qk}^2. In reality σ_{qk}^2 is unattainable, but it is often fruitful (as in Example 2) to use a model where σ_{qk}^2 is a function of an auxiliary variable u_q. Subject knowledge, guesses, or previous survey estimates can be used as planning values of σ_{qk}^2. The next section gives an overview of

schemes for πps sampling, and to connect to the previous sections, we give recent references where πps sampling is combined with GREG estimation.

4. Implementing a πps design

The designs discussed in the previous section all suggest that unequal probability sampling should be used (π_k should be chosen proportional to some measure z.) Hence, we need a sampling scheme that implements $\pi ps(z)$ designs. The use of $\pi ps(z)$ designs has a long history in survey sampling, and it is one way of using auxiliary information in the design stage. However, much of the discussion in the literature focuses on strategies where a $\pi ps(z)$ design is combined with the π estimator. The reason for this is tradition. When a study variable y_q is strictly proportional to z, and $\pi_k = nz_k / \sum_U z_k$, we have $\hat{t}_{y_q\pi} = n_s t_{y_q} / n$. This means that with a random size $\pi ps(z)$ design, $\hat{t}_{y_q\pi}$ will vary due to the variation in sample size only, and for a fixed size design, $\hat{t}_{y_q\pi}$ has no variation at all.

Nevertheless, given a $\pi ps(z)$ design, there is no need to restrict ourselves to the π estimator. Since we have auxiliary information at hand, we can use the GREG estimator. Furthermore, with GREG estimation our loss in efficiency by using a random size design, instead of a fixed size design, is likely to be small (if $E_p(n_s)$ is not too small.) Random size designs like the traditional Poisson πps sampling, or the recently proposed PoMix sampling, are described in Kröger, Särndal and Teikari (1999).

However, statisticians often prefer fixed size designs. They enable control over the sample size and the statistician avoids the task of explaining to clients that the initial sample size, and thereby the cost of the survey, is a random component. Over the years, these circumstances have led to an extensive effort to find a selection scheme for implementing a fixed size $\pi ps(z)$ sampling design. Although many sample selection schemes have been proposed, it has turned out to be hard to devise a fixed size scheme for an arbitrary sample size n that has a number of desirable properties, such as (a) the actual selection of the sample is relatively simple, (b) all first-order inclusion probabilities are strictly proportional to the size variable, (c) the design admits (at least approximately) unbiased estimation of the design variances $V(\hat{t}_{y_q\pi})$ and $V(\hat{t}_{y_qr})$. If we also want to use the technique of permanent random numbers (*PRN*) in the sample selection, which is desirable in large survey organizations, it will be even harder.

Nevertheless, for statisticians preferring fixed size πps sampling, some sampling schemes fulfill most of the requirements above. Relatively new fixed size πps designs are *order-sampling* designs, as Pareto πps and sequential Poisson πps (see Rosén (1997), Saavedra (1995) and Ohlsson (1995) respectively). Fixed size *PoMix* proposed by Kröger, Särndal and Teikari (2000) is another alternative, and comparisons have shown (see Holmberg (2001) and Holmberg and Swensson (2001)), that also *model-based stratified simple random sampling* (mb-STSI) proposed by Wright (1983) is a method that should be considered.

However, which sampling scheme to use is not the issue here, and depending on the situation there are pros and cons for all of them. Here, we merely state that alternatives that approximately fulfill the requirements above exist. Rosén (1997, 2000a, 2000b) and the references therein provide details on the Pareto πps, which is used in the Swedish Crop Yield Survey and Swedish Market Tendency Survey. PoMix sampling is described in the references by Kröger et al. (2000), and Holmberg, and mb-STSI can be studied in Wright as well as in Särndal et. al chapter 12.

5. A numerical comparison of the multi-parameter approaches

In this section, we will give an example on the use of the above multiparameter approaches, and how it is possible to use information collected at the planning stage for further elaboration of and for support in the choice of sampling design.

One of the purposes of this paper is to suggest approaches that might be useful to achieve a high overall efficiency. In a factual survey situation, the success in achieving this depends highly on the quality and structure of the auxiliary information and how the statistician uses the auxiliary information. The usefulness of the approaches in previous sections will therefore vary from case to case. However, we can mimic a real survey situation to give an idea on how they can be applied and how they might work in practice.

In the next section's example, we use a real and easily accessible finite population, (the population of Swedish municipalities MU281 available in appendix B in Särndal et al.) We place ourselves at the planning stage of a multiparameter survey from this population, where we are supplied with auxiliary variables, and where we have certain more or less valid beliefs and guesses about the relationships between our study variables and these auxiliary variables. The chosen relationships are not necessarily the best for this specific survey population, yet chosen to mimic a realistic starting point for a statistician planning a survey with auxiliary information, and to mimic a situation where the involved study variables are thought to have varying relations with the available auxiliary variables. (The latter is common in farm surveys, where some auxiliary variables

that work well for farms specialized on crop might be poor for farms specialized on animals and vice versa.) Hence, the relationships we use suggest a point estimator (a GREG estimator) for each parameter, and they give us ideas of alternative sampling designs where the auxiliary information can be utilized. Altogether, our planning stage conditions give us a variety of alternatives for selecting a sampling design. The statistician subjectively determines many of these conditions, (i.e. through his beliefs about the relationships between the auxiliary variables and the study variables, and his choice of important parameters.) Still, given these conditions, we can compare the design alternatives, as is done below. In the end, the results of such comparisons might lead to a design decision that is good in meeting the overall demands of the survey.

5.1. Planning a multiparameter survey to achieve overall efficiency: An example

We use all the quantitative variables in the MU281 population. As auxiliary variables we have $u_1 = P75$ (1975 population), $u_2 = S82$ (total number of seats in the municipal council 1982) (and a constant $u_{3k} = 1$ for every $k \in U$.)

The six study variables are:

P85	(1985 population)
RMT85	(Revenues from the 1985 municipal taxation)
ME84	(Number of municipal employees 1984)
REV84	(Real estate values according to 1984 assessment)
CS82	(Number of conservative seats in municipal council 1982)
SS82	(Number of social democratic seats in municipal council 1982

We plan to use GREG estimators $\hat{\mathbf{t}}_r = \left(\hat{t}_{y_1 r}, \ldots, \hat{t}_{y_6 r}\right)'$ to estimate $\mathbf{t} = \left(t_{y_1 r}, \ldots, t_{y_6 r}\right)'$, and the parameters are rated as equally important, (i.e. $H_q = 1/6$ for $q = 1, \ldots, 6$.) To assist in the planning of the sampling design, we have some a priori ideas of the relations between the study variables and the auxiliary variables. These are described in table 1, i.e. we believe it is reasonable to apply linear models $\xi_q \ \left(q = 1, \ldots, 6\right)$, according to (1), where σ_{qk}^2 are substituted with 'guesstimates' $\tilde{\sigma}_{qk}^2$.

Table 1. Planning stage assumptions for the relations between the auxiliary variables and the study variables

q	1	2	3	4	5	6
y_q	P85	RMT85	ME84	REV84	CS82	SS82
$\mathbf{x}_q^{'}$	$(1, P75)$	$(1, P75, S82)$	$(1, P75)$	$(1, P75)$	$(1, P75, S82)$	$(1, S82)$
$\tilde{\sigma}_{qk}^2$	$P75^{\tilde{\gamma}_q}$	$P75^{\tilde{\gamma}_q}$	$P75^{\tilde{\gamma}_q}$	$P75^{\tilde{\gamma}_q}$	$S82^{\tilde{\gamma}_q}$	1
$\tilde{\gamma}_q$	1.4	2	2	0.4	1.2	0

In this example, the $\tilde{\sigma}_{qk}^2$:s are different functions of different auxiliary variables, and the constant factors, σ_q^2 (discussed in example 2), are assumed to be 1 for every q. (Knowledge of such factors is important when approach A and approach B are applied, but for approach C and the single parameter approach of section 2, it is not necessary.)

Note that the purpose of the information given in table 1 is not to illustrate some true or even necessarily good model relations for this population. The purpose of the chosen model relations is to reflect what might be a realistic planning stage situation. By this, we mean a situation where the statistician, for each parameter individually, believes that the survey can benefit substantially from using the auxiliary information in the design as well as in the estimator. Furthermore, the six model relations in table 1 illustrate flexible differences in the planned way to use the auxiliary variables, especially with respect to $\tilde{\sigma}_{qk}^2$. For $q = 1,2,3,4$, $\tilde{\sigma}_{qk}^2$ is a function of the auxiliary variable $P75$, for $q = 5$ it is a function of $S82$, while for $q = 6$ it is constant. Obviously, the mixture of different functions for $\tilde{\sigma}_{qk}^2$ will influence the properties of a compromise design. To further clarify, we translate some information in table 1 to more 'well-known' cases. For example, if we insert the values of $\tilde{\sigma}_{qk}$ into equation (6) of section 2, the planning relations of table 1 imply: (i) When it comes to estimating t_{y_3}, then $\tilde{\sigma}_{3k} = u_{1k} = P75_k$, and the planning stage strategy the statistician believes in is a $\pi ps(P75)$ sampling design combined with GREG estimator using $\mathbf{x}_q' = (1, P75)$. (ii) To estimate t_{y_6} he or she suggests a design with equal inclusion probabilities, (i.e. $\dot{\pi}_{6(opt)k} = n/N$ for every $k \in U$), combined with a GREG estimator where $\mathbf{x}_q' = (1, S82)$. A similar kind of reasoning can be used to understand the implications of other planning stage assumptions given in table 1.

We can use the setup from table 1 to create a diagnostic table for the alternative planning stage designs. (Note that the sampling scheme for implementing a design (see section 4) is left open in this section.) By calculating the $\tilde{\sigma}^2_{qk}$ values and applying them to equations (6), (8) and (12), we can determine $\dot{\pi}_{q(opt)k}$, $\dot{\pi}_{(A)k}$ and $\dot{\pi}_{(C)k}$ for $k = 1,\ldots,281$ and $q = 1,\ldots,6$. Then, for the different design alternatives (the different sets of $\dot{\pi}$:s), planning values, $ANV^*_q(\hat{t}_{y_q r})$ and $ANV^*_{q\min}(\hat{t}_{y_q r})$ can be computed from equations (5) and (7). These values can then be studied and used to make a prediction of how the different design alternatives might affect single-estimator and overall precision of the survey. If the information collected from such a prediction also carries over to the implementation of the survey, it is valuable for the final design choice.

Table 2 illustrates a planning stage comparison between the designs considered for a MU281 survey, where $E_p(n_s) = 40$. From the information given in table 1, we have computed $ANV^*_{q\min}(\hat{t}_{y_q r})$ for the six designs considered from the single parameter approach, (here denoted p_i, $i = 1,\ldots,6$.) Then, predicted relative efficiency losses, i.e.,

$$PREL_{p_i,q} = 100\left(\frac{ANV^*_q(\hat{t}_{y_q r})_{p_i}}{ANV^*_{q\min}(\hat{t}_{y_q r})} - 1\right),$$

have been computed for $p_1,\ldots,p_6$, as well as for the compromise designs (p_7 and p_8), that follows from approaches A and C of section 3. For each design alternative, the predictions of the overall, (total), efficiency loss are summarized by the row means, given in the last column. The row means can be interpreted as planning stage predictions of $ANOREL$, i.e.

$$100 \cdot ANOREL^*_{p_i} = \sum_{q=1}^{6} PREL_{p_i,q} / 6.$$

Not surprisingly, table 2 indicates that the smallest $ANOREL^*_{p_i}$, (13%), is obtained for the design following as a result of applying approach C. The design p_1, with π_k 'optimally' chosen for estimating t_{y_1} (i.e. with $\pi_k = \dot{\pi}_{1(opt)k} \propto z_k = u_{1k}^{0.7}$) has the second smallest (17.1%). For design p_1, small (<10%) relative efficiency losses are predicted when t_{y_1}, t_{y_2} and t_{y_3} are to be estimated, but large (>20%) for t_{y_4}, t_{y_5} and t_{y_6}. The designs p_2 and p_3 (which by the way are identical) and the design following from approach A, also predict small losses for t_{y_1}, t_{y_2} and t_{y_3} and large for estimating t_{y_4}, t_{y_5} and t_{y_6}. For the designs p_4, p_5 and p_6, the patterns are reversed. From a multiparameter

perspective, none of the designs $p_1,\ldots,p_7$ seems to be satisfactory as compromise designs.

Table 2. Planning stage relative efficiency losses, $100(\frac{ANV_q^*(\hat{t}_{y_q r})_{p_i}}{ANV_{q\min}^*(\hat{t}_{y_q r})}-1)$, for eight alternative sampling designs, ($E_p(n_s)=40$) when estimating six population totals of MU281. (Boldface numbers show the largest efficiency loss for each design.)

Design Approach	Parameters						
	t_{y_1}	t_{y_2}	t_{y_3}	t_{y_4}	t_{y_5}	t_{y_6}	$ANOREL_{p_i}^*$
p_1: 'Optimal' for t_{y_1}	0	8.6	8.6	20.1	25.8	**39.4**	17.1
p_2: 'Optimal' for t_{y_2}	8.0	0	0	57.9	68.5	**93.4**	39.7
p_3: 'Optimal' for t_{y_3}	8.0	0	0	57.9	68.5	**93.4**	39.7
p_4: 'Optimal' for t_{y_4}	23.3	**72.2**	**72.2**	0	0.7	2.8	28.5
p_5: 'Optimal' for t_{y_5}	29.2	**83.5**	**83.5**	0.7	0	1.8	33.1
p_6: 'Optimal' for t_{y_6}	49.1	**125.7**	**125.7**	3.0	1.9	0	50.9
p_7: Approach A	3.6	1.1	1.1	38.9	45.7	**63.7**	25.6
p_8: Approach C	2.9	17.4	17.4	8.9	11.6	**19.6**	13.0
$ANV_{q\min}^*(\hat{t}_{y_q r})$	119192	845420	845420	5428.4	170311	1693.0	

Remark 4 Concerning approach B, our data does not permit a fair comparison with the other approaches, and we suspect that in most practical situations, the information needed at the planning stage is too demanding. However, sometimes a planning value for $\mathbf{B}_q$*, say* $\dot{\mathbf{B}}_q$*, might be available, and then*

$$\dot{t}_{y_q}^2 = \left(\sum_U \mathbf{x}'_{qk}\dot{\mathbf{B}}_q\right)^2 \tag{17}$$

could be used as a planning value for $t_{y_q}^2$.

5.2. Design comparisons on population data

The results in table 2 give rough guidelines of the properties of the considered designs. For any real finite population, the model assumptions made at the planning stage will deviate more or less from factual conditions. Therefore, actual calculation of estimator variances from our population, will give valuable information on what would have happened, if we had implemented the planning stage ideas of table 1. It will also give indications on to what extent the predicted design properties, as those of table 2, are transferable and valid for the actual sample survey.

For all our six parameters $t_{y_1},\ldots,t_{y_6}$, we consider GREG estimators, $\hat{t}_{y_1 r},\ldots,\hat{t}_{y_6 r}$ (see equation (2)), using $\mathbf{x}'_{qk}$ and $c_{qk}=\tilde{\sigma}_{qk}$ from table 1. An easy way to compare the alternative designs of table 2, is to calculate the estimator variances under Poisson sampling. For Poisson sampling, the Taylor expansion variance of equation (4) is $V_{T_{(PO)}}(\hat{t}_{y_q r})=\sum_U\left(\pi_k^{-1}-1\right)E_{qk}^2$ and we calculated $V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}$ for $q=1,\ldots,6$ and all considered designs p_i, $\left(i=1,\ldots,8\right)$. Table 3 is based on the results from those variance calculations.

To simplify comparisons, table 3 has the same structure as table 2. Thus, we show results for all the eight alternative sampling designs considered at the planning stage. For each parameter, we determine the smallest estimator variance V^* over all the considered designs, i.e. $V_q^*(\hat{t}_{y_q r})=\min_{i=1}^{8} V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}$ is calculated for $q=1,\ldots,6$. Given the information we used at the planning stage, the values of $V_q^*(\hat{t}_{y_q r})$ will then represent the 'best' result we might obtain for every parameter separately. Dividing every $V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}$ by $V_q^*(\hat{t}_{y_q r})$ parameter-by-parameter, we get a measure comparable to the predicted relative efficiency losses, $PREL_{p_i,q}$, shown in table 2.

The general pattern of the relative efficiency losses in table 3 is the same as in table 2. Hence, given the selected GREG estimators, table 2 gives a good image of the relative efficiency losses for the considered designs under Poisson sampling. Parameter by parameter, the design based on approach C is never the best alternative, but in an overall sense it is the most efficient, with (13.1%) mean efficiency loss. Therefore, it can be argued that it also is the best compromise design, followed by design p_1, just as in table 2. One notable difference between the planning values of table 2 and the population values in table 3, is that design p_4 is slightly better than p_5 for estimating t_{y_5}. This is likely to happen in reality as well, the planning values are just guesses, more or less accurate, and $ANV_q^*(\hat{t}_{y_q r})$ and $V_{T_q}^*(\hat{t}_{y_q r})$ are different.

Table 3. Estimated relative efficiency losses, $100(\frac{V_{T_{(PO)q}}(\hat{t}_{y_q r})_{p_i}}{V_q^*(\hat{t}_{y_q r})} - 1)$, for eight alternative Poisson sampling designs, ($E_p(n_s) = 40$), when estimating six population totals of MU281. (Boldface numbers show the largest efficiency loss for each design.)

Design Approach	Parameters						
	t_{y_1}	t_{y_2}	t_{y_3}	t_{y_4}	t_{y_5}	t_{y_6}	$OREL(\%)$
p_1 : 'Optimal' for t_{y_1}	0	6.9	6.7	21.2	18.7	**37.8**	15.2
p_2 : 'Optimal' for t	8.9	0	0	59.2	54.3	**90.1**	35.4
p_3 : 'Optimal' for t_{y_3}	8.9	0	0	59.2	54.3	**90.1**	35.4
p_4 : 'Optimal' for t	19.6	**58.7**	56.3	0	0	2.7	22.9
p_5 : 'Optimal' for t	26.8	**65.8**	62.3	1.8	2.1	1.1	26.6
p_6 : 'Optimal' for t	42.4	**101.1**	96.2	2.3	2.9	0	40.8
p_7 : Approach A	4.7	1.7	1.6	41.4	36.1	**60.7**	24.4
p_8 : Approach C	3.8	17.9	17.9	10.1	9.6	**19.1**	13.1
$V_q^*(\hat{t}_{y_q r})$	6810	895719	6.36E7	1.75E9	20156	37271	

Remark 5 Table 3 is based on Poisson sampling estimator variances. Similar tables with similar results can be computed for fixed size designs (e.g. using one of the sampling schemes mentioned in section 4.) However, since there will be more approximations involved, the results might be somewhat less reliable.

Remark 6 The designs p_2 (or p_3) and p_6 in tables 2-3 are in a sense traditional textbook designs. In p_6, all inclusion probabilities are equal as in simple random sampling or (as for the data in table 3) Bernoulli sampling. p_2 is traditional since the inclusion probabilities are proportional to the most useful auxiliary variable P75, i.e. ($\pi_k \propto u_{1k}$). An untransformed auxiliary variable is often used

as size measure in πps designs. It should be noted from tables 2-3 that these latter designs are the worst in overall efficiency terms. Therefore, from a multiparameter perspective, they cannot be recommended in situations similar to the one described.

If the effects of using a design such as p_8 not are entirely satisfactory, then a more elaborate compromise design is possible by applying approach D, given in the next section.

6. Approach D: Minimizing the weighted sum of relative efficiency losses under restrictions.

This section outlines yet another multiparameter alternative. Here, we choose to extend approach C, although similar reasoning can be applied for approaches A and B as well.

The statistician may consider this approach from start, but for two reasons we suggest that the implications of approach C (which gives a minimum ANOREL) are examined first. Firstly, realistic restrictions are more easily found after studying results from approach C, and secondly, approach C will provide useful numerical knowledge for the calculations in approach D.

If we once again study table 2, and if we regard approach C as the approach most suitable for our goals, we observe that, although approach C implies the smallest overall efficiency loss, these losses vary between the parameter estimates. For example in table 2 we have more than 15% efficiency losses for the parameters t_{y_2}, t_{y_3}, and t_{y_6}, while the others have smaller values.

We might want a design where the sum, $(ANOREL^*_{p_8})$, still is small, but where certain restrictions on the individual variances are fulfilled, so that the efficiency loss compared to the 'optimal' variance does not exceed, say c_q. Hence, we can formulate the non-linear optimization problem below, where the specified restrictions now make the importance weights H_q redundant.

Problem 7 *Minimize the objective function*

$$f(\boldsymbol{\pi}) = \sum_{q=1}^{Q} \sum_U \left(\pi_k^{-1} - 1\right) \frac{\sigma_{qk}^2}{\sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2}$$

where $\sigma_{qk}^2 \geq 0$ *and* $\sum_U \left(\pi_{q(opt)k}^{-1} - 1\right)\sigma_{qk}^2 > 0$ *under the* $2N + Q + 1$ *restrictions*

$$0 < \pi_k \leq 1 \qquad k = 1, \ldots, N$$

$$g_0(\boldsymbol{\pi}) = \sum_U \pi_k - n = 0$$

$$g_q(\boldsymbol{\pi}) = \sum_U \left(\pi_k^{-1} - 1\right) \frac{\sigma_{qk}^2}{ANV_q} \leq c_q \quad q = 1, \ldots, Q$$

The number of restrictions and the mixing between linear and non-linear restrictions, as well as equality and inequality restrictions complicate the problem. Therefore, it is hard to find useful analytical solutions. However, the Karush-Kuhn-Tucker conditions from optimization theory apply (see e.g. Lundgren, Rönnqvist and Värbrand (2001) or Fiacco and McCormick (1990)) and if the Q $g_q(\boldsymbol{\pi})$-restrictions are not too strictly set, solutions can be obtained with non-linear programming algorithms. Hence, with some effort the flexibility in choosing a compromise design can be increased.

7. Conclusions

Planning a multipurpose survey with several important parameters is not a straightforward task. In this paper, we have presented some potentially useful approaches when auxiliary information is available.

By adapting a multiparameter perspective already at the planning stage, we illustrate that compared to a single parameter approach; significant improvements on the overall precision of a survey are possible. If we plan for the possibility of using the auxiliary information in the sampling stage as well as in the estimation stage, we can, by conditioning on an efficient estimator such as the GREG estimator and focusing on the design choice, construct diagnostic tables such as the one exemplified by table 2 of section 5. This can give us valuable information to compare the properties of the design that best fulfills our goals. Since the final survey plan depends on the overall objectives, we cannot give absolute recommendations on the planning of a multiparameter survey. However, approach C of this paper seems to be a useful approach. The multiparameter perspective used in that approach takes into consideration 'optimal' results for the single parameter case, and by minimizing a relative measure it seems as if the approach

has an overall robustness. That is, the loss in efficiency compared to best possible single parameter solutions will not be extremely high for any parameter estimate. Furthermore, if the solution from applying approach C not is satisfactory, a certain amount of flexibility in the design choice is possible. Under certain regularity conditions, non-linear programming can be used to construct a design that fulfills certain precision requirements.

A detailed discussion of the sampling schemes that can be used to implement a πps sample is beyond the scope of this text. However, there has been progress in that area in recent years, especially concerning fixed size sampling designs. Detailed information is provided in the references on πps sampling in section 4. Finally, a crucial issue for applying the results in this paper is to have good planning values of σ_{qk}^2. In practice, statisticians often seem to use the approximation $\sigma_{qk}^2 \approx u_{qk}^{\tilde{\gamma}_q}$, where $\tilde{\gamma}_q$ is an estimated or guessed value of a parameter γ_q. (Harvey (1976) describes how ML-estimates of γ_q can be obtained, and in finite populations, γ_q often lies in the interval $(0,2)$ or according to Brewer (1963) in the narrower interval $(1,2)$.) Results from Rosén (2000a) and Holmberg & Swensson indicate, that the positive effects of having good $\tilde{\gamma}_q$ values (or $\tilde{\sigma}_{qk}^2$ values) for the design, can be substantial. In addition, from the example in the present paper we also note that unreflectively chosen values, e.g. choosing a design implicitly based on $\tilde{\gamma}_q = 2$, can have a large negative effect in a multiparameter perspective (see designs p_2 and p_3 of tables 2-3.) Hence, more attention (than we believe is the case today) should be paid on finding good planning values for σ_{qk}^2. Especially in surveys repeated over time, such attention could be a relatively cheap way to improve the survey quality.

Acknowledgements:

The author thanks Prof. Bengt Swensson, Dr. Martin Axelson and the referee for their valuable comments on earlier versions of this article.

REFERENCES

BELLHOUSE, D. R., (1984). A review of optimal designs in survey sampling. *Canadian Journal of Statistics*, 12, pp 53-65.

BREWER, K. R. W., (1963). Ratio Estimation and Finite population: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* 5, 93-105.

CASSADY, R. J. and VALIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.

CASSEL, C. M., SÄRNDAL, C-E. and WRETMAN, J., (1976). Some results on generalized difference estimators and generalized regression estimators for finite populations. *Biometrika* 63, 615-620.

CASSEL, C. M., SÄRNDAL, C-E. and WRETMAN, J., (1977). *Foundations of Inference in Survey Sampling.* Wiley & Sons, New York

DEVILLE, J.-C. and SÄRNDAL, C-E., (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, 87, 376-382.

ESTEVAO, V. and SÄRNDAL, C.E., (2000). A Functional Form Approach to Calibration. *Journal of Official Statistics*, 16, No. 4, 379--399.

FIACCO, A.V. and MCCORMICK, G.P., 1990. *Nonlinear Programming -- Sequential Unconstrained Minimization Techniques*, Classics in applied mathematics, SIAM.

HARVEY, A.C., (1976). Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrika*, 44, No. 3, 461-465

HOLMBERG, A., (2001). On the Choice of Strategy in Unequal Probability Sampling. *Proceedings of the section on Survey research Methods, Joint Statistical Meetings*, Atlanta 2001, American Statistical Association, CD-ROM.

HOLMBERG, A. and SWENSSON, B., (2001). On Pareto πps sampling: Reflections on unequal probability sampling strategies. *Theory of Stochastic Processes*, 7(23), No. 1-2 (2001) 142-155.

ISAKI, C. T. and FULLER, W. A., (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, 77, 89-96.

KRÖGER, H., SÄRNDAL, C-E. and TEIKARI, P., (1999). Poisson Mixture Sampling: A family of designs for Coordinated Selection Using Permanent Random Numbers, *Survey Methodology*, 25, No 1, 3-11.

KRÖGER, H., SÄRNDAL, C-E. and TEIKARI, P., (2000). Poisson Mixture Sampling Combined with Order Sampling: a Novel use of the Permanent Random Number Technique. Manuscript submitted for publication (date 00/08/30). (Forthcoming in Journal of Official Statistics with the title Poisson Mixture Sampling Combined with Order Sampling)

LUNDGREN, J., RÖNNQVIST, M., and VÄRBRAND, P., (2001). *Linjär och icke-linjär optimering,* Studentlitteratur Lund.

LUNDSTRÖM, S. and SÄRNDAL, C-E., (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics,* 15, 305-327.

MONTANARI, G.E., (1987). Post-sampling efficient QR-prediction in large-scale surveys. *International Statistical Review,* 55, 191-202.

MONTANARI, G.E., (1998). On Regression Estimation of Finite Population Means. *Survey Methodology,* 24, No 1, 69-77.

OHLSSON, E., (1995). Sequential Poisson Sampling. Research Report from Institute of Actuarial Mathematics and Mathematical Statistics at Stockholm University.

RAO, J. N. K., (1979). Optimization in the design of sample surveys. In: J.S. Rustagi (ed.), *Optimization methods in Statistics: Proceedings of an International Conference*. New York, Academic Press, pp 419-434

RAO, J. N. K., (1992). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. *Proc. Workshop on Uses of Auxiliary Information in Surveys,* Statistics Sweden.

RAO, J. N. K., (1994). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. *Journal of Official Statistics,* 10, 153--165.

RAO, J. N. K., (1997). Developments in sample survey theory: an appraisal. *Canadian Journal of Statistics,* 25, 1-21.

ROSÉN, B., (1997). On sampling with Probability Proportional to Size, *Journal of Statistical Planning and Inference,* 62, 159-191.

ROSÉN, B., (2000a). Generalized Regression Estimation and Pareto πps R & D report 2000:5 Statistics Sweden.

ROSÉN, B., (2000b). A User's Guide to Pareto πps Sampling, R & D report 2000:6 Statistics Sweden.

SAAVEDRA, P., (1995). Fixed Sample Size PPS Approximations with a Permanent Random Number. *Proceedings of the section on Survey research Methods Joint Statistical Meetings,* American Statistical Association, *697-700.*

SÄRNDAL, C-E., SWENSSON, B. and WRETMAN, J., (1992). *Model Assisted Survey Sampling.* Springer, New York.

WRIGHT, R. L., (1983). Finite Population Sampling with Multivariate Auxiliary Information, *Journal of the American Statistical Association,* 78, 879-884.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 995–1011

WEB TOOLS IN TEACHING AND LEARNING OF SURVEY SAMPLING: THE VLISS APPLICATION

Vesa Kiviniemi and Risto Lehtonen[1]

ABSTRACT

We describe in this article a web application prepared for university-level teaching and learning of certain aspects of Survey sampling, a specific area of Statistical science. Survey sampling covers such areas as sampling and estimation for finite populations, edit and imputation, and the design and analysis of complex surveys. Survey sampling methodologies are routinely used in official statistics production in many countries. The methods also are increasingly used in empirical research in different disciplines. Teaching of survey sampling usually takes place in university departments of statistics. Our web application has been built in University of Jyväskylä, and is entitled as "VliSS-Virtual laboratory in Survey Sampling". The core of the current VliSS contains materials supporting the use of a recent textbook in university level teaching of Survey sampling. The application can be accessed via Internet. Our main goal in building the application has been to take advantage of the special features of Internet, that is, dynamic and interactive features, updating possibilities and wide access. We have included in the application not only fully worked pedagogical examples and dynamic graphics extending the textbook materials but also options to download data files and program codes for personal interactive training. We will demonstrate some of the interactive and dynamic aspects of the VliSS in this paper. In addition, we discuss to some extent the possible problems in building of a web-based learning environment and solutions chosen for our application. Further plans on research and development of the VliSS application are briefly discussed. The properties of the application will be investigated empirically using data produced by the application, and certain item analysis tools such as the Rasch model and its extensions will be used.

Key words and phrases: Web-based teaching and learning, Survey sampling, Item analysis, Internet

[1] University of Jyväskylä, Department of Mathematics and Statistics, P.O.Box 35, FIN-40014 University of Jyväskylä, Finland. vesa.kiviniemi@jyu.fi risto.lehtonen@maths.jyu.fi

1. Introduction

During the past decade the change towards an information society has been effectively implemented throughout the industrialized world with the aid of information technology (Webster 1995; Isomäki 2002). Development of information technology has also invoked educators in universities and elsewhere to modify prevailing teaching methods. At the same time calls for restructuring the way students learn come from a variety of institutions. Web-based learning has been one possibility to restructure the way of learning. (Brandon 1996; Dodge 1995.) These approaches include for example pedagogical introduction to the subject and guidance for the use of a web-based learning tool.

A "hybrid" solution constituting of a combination of a paper-printed textbook and a web application that supports the use of the textbook in training provides a challenging option. Our web application "VliSS-Virtual laboratory in Survey Sampling" has been prepared to support the use of the textbook of Lehtonen and Pahkinen, "Practical Methods for Design and Analysis of Complex Surveys" (John Wiley & Sons, 1996), in university-level teaching of basic and more advanced Survey sampling. Although the VliSS interactively supports the use of the textbook, it can also be used to some extent for learning purposes without an access to the textbook. We however believe that at least at the current stage of educational practices and cultures, a paper-printed textbook still provides a strong educational tool.

The VliSS application can be accessed via Internet. Our main aim in building the application has been to take advantage of the special features of Internet, that is, dynamic and interactive features, updating possibilities and wide access. We have included in the application not only fully worked pedagogical examples and dynamic graphics extending the textbook materials but also options to download data files and program codes for personal interactive training,

A link between the textbook and the web application is provided by the so-called Training Keys. Each Training Key refers to a specific page of the printed book and when activated, opens an access to further training and practical application of the methods discussed in the respective page(s) of the textbook. The textbook itself is not included in the web application; only the list of contents of the book supplemented with Training Keys are visible for the user.

The current trial version of the VliSS application covers to some extent the following aspects of basic and advanced Survey sampling and analysis: Finite population sampling and estimation, approximative variance estimation in complex surveys, and design-based multivariate survey analysis. In each area, interactive training modules have been implemented covering options for downloading data files and program codes. Future plans include additional options such as Student's Corner, Teacher's Corner, Help Desk, Links and FAQ pages.

From a technical point of view, The VliSS system is designed by using HTML, JavaScript and SAS programming languages so that the coding supports Internet Explorer 4 and Netscape 6 or higher browsers.

The paper is organized as follows. In Section 2 we discuss some aspects of web-based learning and designing of web-based learning environments. In Section 3, the VliSS application is introduced in more detail and some of the dynamic and interactive solutions are demonstrated. In Section 4, plans for research and development of the VliSS application are introduced. Conclusions are presented in Section 5.

Access to the trial version of the VliSS is available in the address http://www.stat.jyu.fi/mpss/VLISS (Web reference 1). The authors will appreciate user feedback.

2. Web-based learning and learning environments

Education is receiving increasing pressure from changing global economic circumstances and complex societal needs. Calls for restructuring the way students learn come from a variety of institutions. Educators agree that we must help students learn to solve problems and think independently (Bransford et al. 1988). The challenge for educators is to develop educational strategies that teach content in ways that also teach thinking and problem-solving skills (Bransford et al. 1988). Recent development of information and communication technologies has provided new technical possibilities to build information systems for various purposes (Isomäki 2002). In education, web tools have offered a new solution to apply such strategies by providing a rich environment for active learning, since web tools are believed to approach education more in a learner-focused way than in a teacher-focused way (Brandon 1996). This refers to a constructivistic approach of teaching and learning (see e.g. Jonassen et al. 1995).

Web-based learning environments aim at providing an interactive tool in which students can work independently with the possible assistance of tutor or web master. Also an advantage of web-based learning is that it supports the "Learning by doing" ideology (see Dewey 1951), which is highly appreciated in nowadays pedagogy. Web-based teaching and learning seems to support many feasible properties of modern education theories. But, on the other hand, building, maintenance and run of such tools can be demanding and time consuming from the designer's and instructor's point of view. This might hold especially for web-based on-line course applications.

Web-based learning environments can be technically relatively easy to build. However, the number of technical possibilities is enormous. The environment can include in example FAQ files, software archives, links, news groups, reference materials, testing, tutorials, evaluation, feedback database, etc. (Brandon 1996, Dodge 1995, Web reference 2.) Obviously, an appropriate interactive use of the various components requires a properly structured environment (Brandon 1996,

Web Reference 2). Tools for designing the environment include for example such languages or applications as HTML, DHTML, Java Script, JAVA, MOOs (Multi-user object-orientated space), Real Audio, VRML (Virtual Reality Modelling Language). Each week there appears new sources of information and tools that have the potential to enhance learning for all of us. (see for example Web reference 3, Web reference 4, Web reference 5.) A challenge for designers is also to build systems that better support human behaviour (Isomäki 2002).

Many instructions and procedures how to design a web-based learning environment can be found in the web (e.g. Web reference 6, Web reference 7 and Web reference 8, Web reference 9, Web reference 10) and many examples of such environments are readily available in the web (see Web reference 11). Still the problem lies in the fact that although technology has provided more and more capable graphic and interactive tools for designing web-based learning environments, very limited educational or statistical research has been published on how to evaluate and develop environments that would meet the specific demands of different disciplines and effectively serve the purposes of different types of users.

In developing the VliSS application, our purpose has not been to introduce the latest technical development, for example DHTML technologies, ASP (active server page) or JSP (java server page) solutions or virtual reality development. The VliSS system is designed to serve learning purposes in a specific discipline and for a specific audience, and to be a reliable source of information in relevant areas. Survey sampling methodologies are routinely used in official statistics production, and the methods are increasingly used in empirical research in different disciplines. Teaching of survey sampling usually takes place in university departments of statistics. The potential audience of the VliSS application thus covers university students, teachers and instructors (especially in Statistical science), researchers in various disciplines dealing with empirical research, and junior and senior statisticians working in Official statistics, research institutes and business firms. Our aim in developing the Vliss application has been to try to be responsiveness to the variety of needs of this audience. In the future, the development of the VliSS will be based on a carefully planned statistical evaluation of the system. This will be discussed in more detail in Section 4.

The core of the VliSS is the concept of “Training Key” making a bridge between the printed textbook and the web materials. The main pedagogical idea in a Training Key is the following: Starting from a problem introduced in the textbook and worked out in the web pages of the Training Key, the treatment of the problem is extended first by using the “learning by doing” method. A more demanding task is proposed for the user to be worked out under the guidance of the application. When these phases are completed, the treatment of the problem is further extended with providing the user with an option for personal interactive analysis. Downloadable data files and program codes are offered for this purpose.

In interactive analysis, the user is also allowed – and encouraged – to modify the setting for one's personal needs.

3. Training of Survey sampling in the VliSS application

In this section, the VliSS application is described in more detail, and two Training Keys are demonstrated and discussed. In the first Training Key, providing an example of descriptive survey sampling, Monte Carlo simulation techniques are used to examine empirically the bias and accuracy properties of certain estimators of a population total. Our second example discusses design-based multivariate modelling in a complex analytical survey. Displays and graphs have been produced by the current VliSS application.

3.1. Description of the VliSS system

Web materials included in the VliSS system support the use, and extend the materials, of the textbook "Practical Methods for Design and Analysis of Complex Surveys", written by Professors Risto Lehtonen and Erkki Pahkinen, and published by John Wiley & Sons in 1996. Training sessions (Training Keys in the VliSS terminology) in the current trial version of the VliSS include the following:

- Simple random sampling and design effect: Analysing a simple random sample (SRS) drawn without replacement.
- Use of auxiliary information in sampling and estimation: PPS sampling (sampling with probabilities proportional to size) and regression estimation for an SRS sample.
- Approximative variance estimation of non-linear statistics in complex surveys: Linearization method, Jackknife and bootstrap techniques for a stratified paired clusters design.
- Design-based multivariate survey analysis: Logistic analysis of variance (ANOVA) and the analysis of covariance (ANCOVA) for data collected by stratified one-stage and two-stage cluster sampling.

Each Training Key includes a fully worked case study covering data specification, statistical analysis (including access to data and program codes used) and display as well as interpretation of results. Graphical presentations and Monte Carlo simulation techniques are used. For further training, an option for interactive analysis is available. Our current version relies on SAS and SUDAAN software products (see Web reference 12; Web reference 13).

The main page of VliSS includes three separately functioning interactive frames (Figure 1). The menu bar in the left-hand frame includes the list of contents of the textbook supplemented with Training Keys. This frame also provides the link between the textbook and the web application. The two others are "General Information" and "Alphabetical Index of Concepts" frames. The

latter works interactively with the main frame providing more detailed information on concepts appearing in the text of the main frame. In Figure 1, the system is displayed when Training Key 106 is accessed.

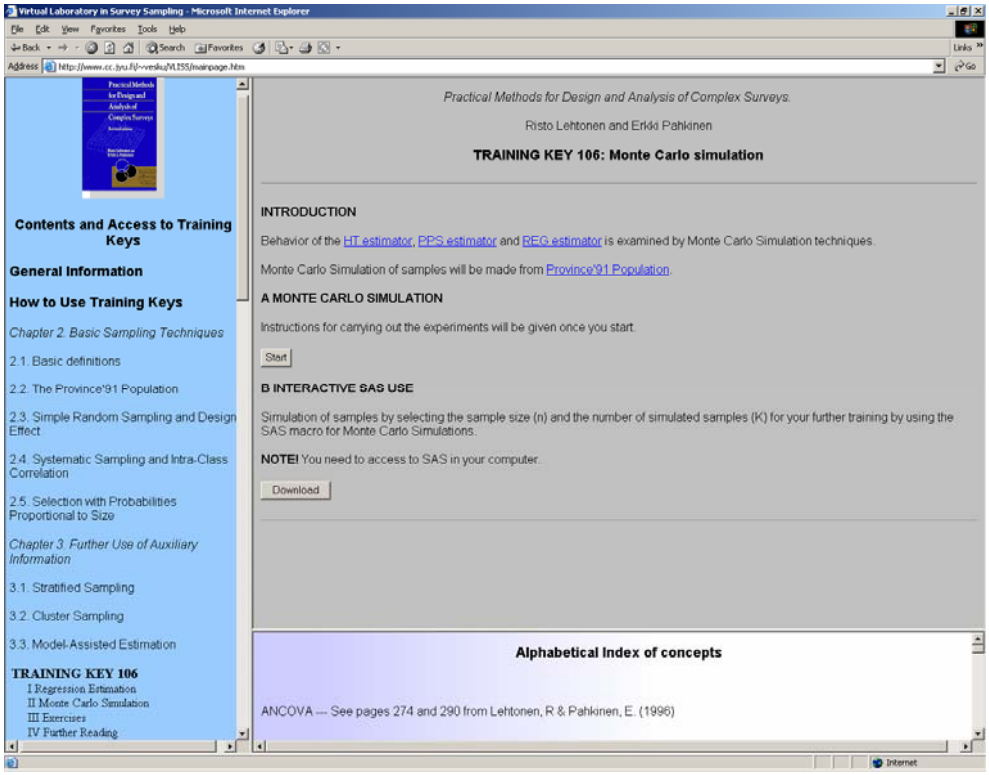


Figure 1. Display of the VliSS when Training Key 106 is accessed

In the VliSS system, each Training Key includes a figure, which refers to the respective page number of the book (i.e. Training Key 106 refers to page 106 of the book). By activating a Training Key the user will enter to the start page of the corresponding training session (worked example, guided demonstration, interactive exercise, more advanced interactive use). In each Training Key, detailed instructions on the use and interpretation of the materials are provided.

The general structure of a Training Key is as follows: First, the relevant materials of the textbook are worked out (problem setting, data display, computation, interpretation). This is followed by an extended example (going beyond the textbook materials) or a simulation exercise, which are worked out under guidance of the application. Finally, an option for interactive further training is offered by downloading the data file and program code on personal PC. The user can then modify the code for personal purposes.

The VliSS application was designed to support Internet Explorer 4.0 and Netscape 6.0 or higher browsers. The system was coded by using HTML, JavaScript, SAS and SUDAAN programming languages. Certain functions programmed by the PERL language are implemented tentatively.

3.2. Training Key: Bias and accuracy of estimators of population total

In Training Key 106 (referring to page 106 of Lehtonen and Pahkinen 1996), Monte Carlo simulation techniques are used for an examination of bias and accuracy of certain estimators of a population total. The aim is to demonstrate the effect of the incorporation of population-level auxiliary information in a sampling design or, alternatively, in an estimation procedure for a given sampling design.

A combination of a sampling design and an estimation procedure is called a strategy. The strategies in this Training Key are (1) SRSWOR-HT, simple random sampling (SRS) without replacement with a standard SRS (or Horvitz-Thompson, HT) estimation procedure (no auxiliary information is used), (2) PPS-SYS-STR, stratified systematic PPS sampling with a standard Horvitz-Thompson estimation procedure (using information of sizes of population elements incorporated in inclusion probabilities), and (3) SRSWOR-REG, regression estimation for a given SRS sample (using auxiliary information in the estimation procedure) (Lehtonen and Pahkinen 1996). The small register-based population consists of N=32 municipalities of a county in Finland. The parameter to be estimated is the total number of unemployed in the county. The sample size is n=8 municipalities. In the SRSWOR-HT strategy, no auxiliary information is used. In the PPS-SYS-STR strategy, auxiliary information (the number of households according to Population Census 1985) is used in the *sampling* phase. In the SRSWOR-REG strategy, similar auxiliary data are used in the *estimation* phase.

The goal in this Training Key is to demonstrate the effect of randomness induced by the sampling design to the behaviour of the estimators in the strategies considered. Such properties as unbiasedness, accuracy and finite population consistency can be illustrated using this Training Key. The introductory example, based on a single sample drawn from the population, is presented and discussed in the textbook. When activating Training Key 106 the user can start working with the extended materials on this estimation problem. To examine empirically the properties (bias and accuracy) of the three different strategies in more detail, the user is guided to draw several independent samples from the population with the given sampling design. The measures used to examine the relative behaviour of the estimators are the Monte Carlo mean and standard deviation, bias, absolute relative bias (ARB) and root mean squared error (RMSE), calculated on the basis of the distributions of the estimators generated by the Monte Carlo experiments. The distributions of the estimators can also be examined graphically. Finally, a summary table can be displayed containing the results from all simulation experiments.

All these tasks are carried out by Training Key 106 section "Monte Carlo Simulation". The user is allowed to select different numbers of samples drawn from the population in order to examine the progress in the estimates and, especially, in the graphical displays. Obviously, when drawing 1,000 samples of size of eight municipalities, the picture of the relative behaviour of the estimators becomes to be quite stable (Figures 2 to 4). All three estimators appear to be essentially unbiased, as expected. The effectiveness of the strategies that incorporate auxiliary information either in sampling phase or in estimation phase appears to be dramatic, when compared to the pure SRS strategy. This is due to the fact that the distribution of the study variable in the population is very skewed, as happens often in real world applications.

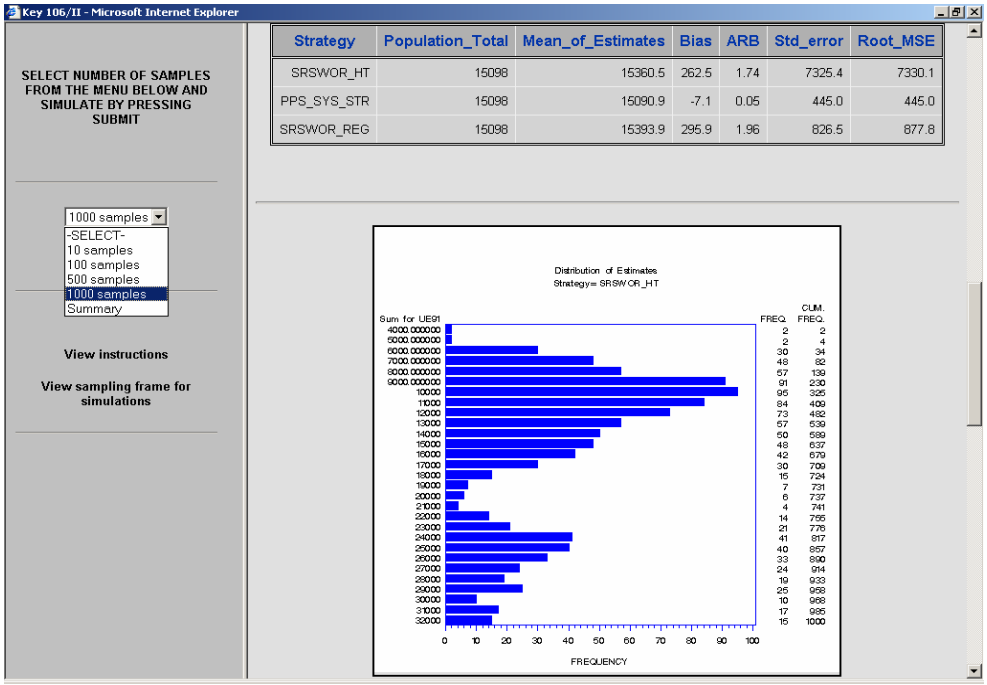


Figure 2. Training Key 106. Descriptive statistics of the simulation experiments for the different strategies, and sampling distribution of the total estimator based on the SRSWOR-HT strategy.

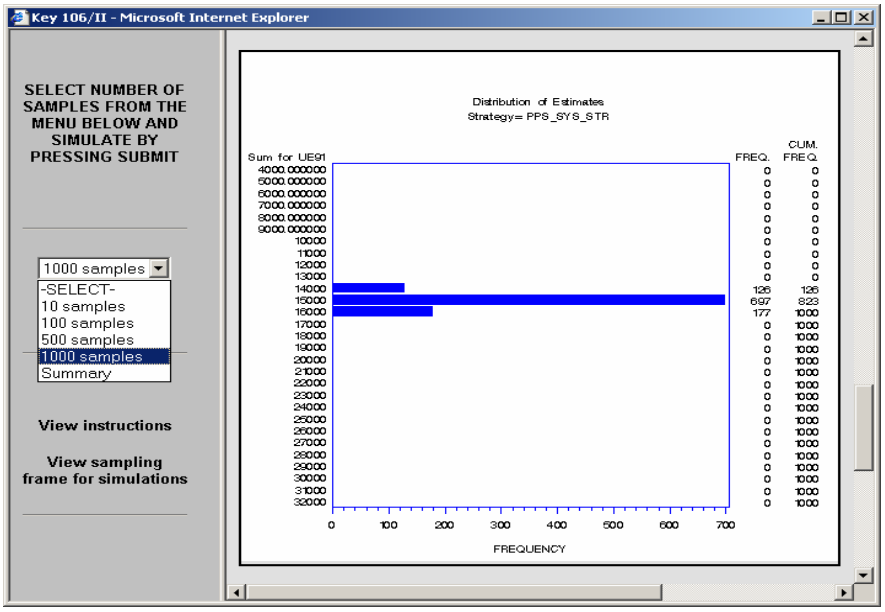


Figure 3. Training Key 106. Sampling distribution of the total estimator based on the PPS-SYS-STR strategy.

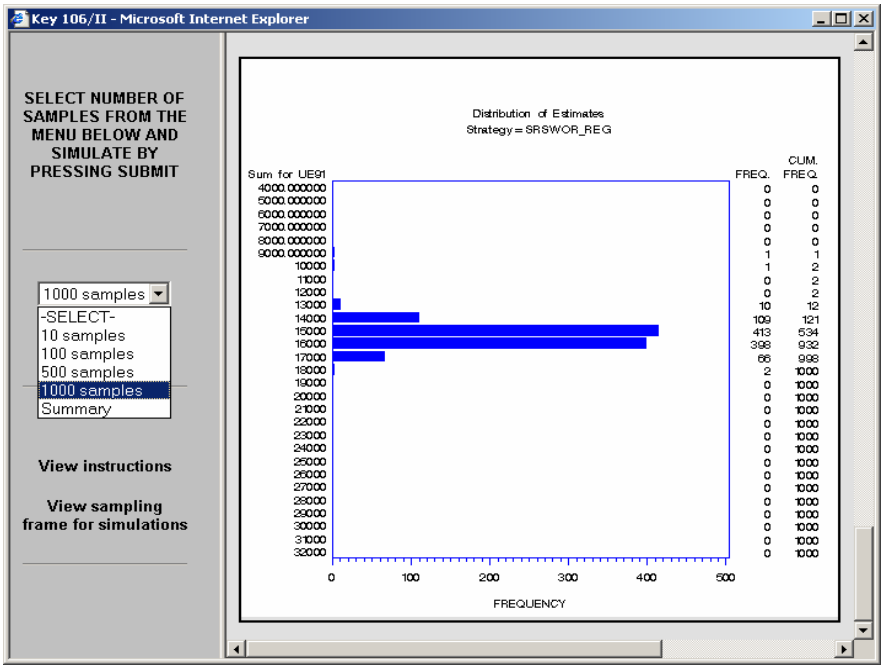


Figure 4. Training Key 106. Sampling distribution of the total estimator based on the SRSWOR-REG strategy.

For further training, Training Key 106 provides an option for interactive analysis where the user can download the population frame data set and the program code on a personal PC. The user can then select the sample size and the number of simulated samples, and study more closely the behaviour of the estimators in the three strategies.

3.3. Training Key: Design-based multivariate analysis

Training Key 262 (referring to page 262 of Lehtonen and Pahkinen 1996) provides an access to design-based multivariate modelling of a binary response variable in a complex survey. The main aim of this exercise is to familiarize the user with stepwise model building in a situation where all predictor variables are categorical. Special interests are in demonstrating the role of interaction terms in a logit ANOVA model. The effect of the removal of an interaction term is examined by graphical presentations. The analysis is essentially design based due to the complex structure of the data set used. The data for this Training Key is taken from a real survey (over 7800 observations), which is based on stratified cluster sampling. The phenomenon under study (psychic strain) appears to be positively intra-cluster correlated, as is indicated for example by the design effect estimates of estimated model coefficients, which tend to be greater than one (Figure 5).

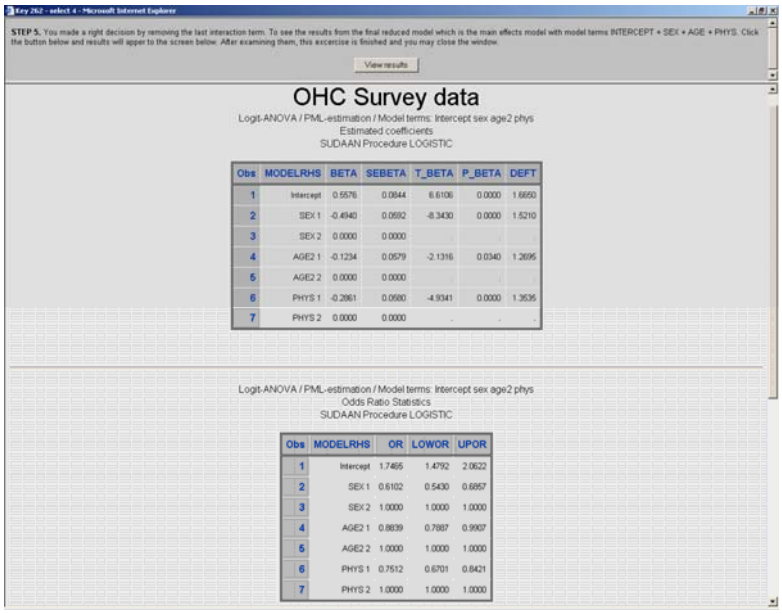


Figure 5. Training Key 262: Step-wise building of a logit ANOVA model. Display of the final main effects model. The estimated coefficients, their standard errors, t-test values, p-values and design effects are displayed. In the lower table, estimated odds ratio statistics are displayed.

Also in this case, the basic example is introduced in the textbook. The web materials in Training Key 262 provide an extended treatment of the modelling problem. The user is encouraged to work out a step-by-step model building procedure. The model selection procedure begins from the saturated model (which includes all the main effects and interaction terms). The aim is to end up with a reduced model that is parsimonious and fits reasonably well. In all phases, odds ratio statistics are estimated and fitted proportions are calculated and displayed graphically (Figures 5 and 6). The system is built to guide interactively the navigation of the user. To make progress, the user must make sensible selections on the model terms to be removed. Once the exercise is completed, the final reduced model can be used for the interpretation of the relationships of the predictors with the response variable.

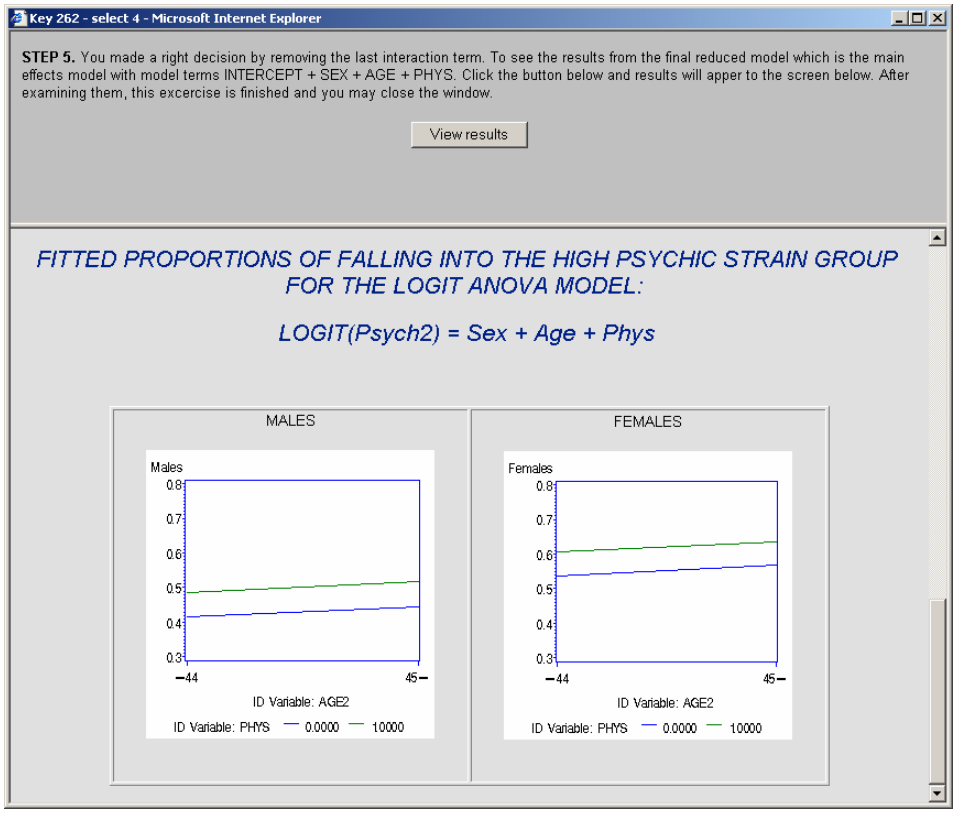


Figure 6. Training Key 262: Step-wise building of a logit ANOVA model. Display of estimated proportions given by the final main effects model.

For further training, Training Key 262 again provides an option for interactive analysis where the user can download the data file and program code for a more detailed examination of the modelling problem.

4. Research and development of the VliSS application

We feel that successful use and further development of a web application requires user feedback and follow-up research on the various aspects of the application The functions and usability of the VliSS will be examined empirically. The follow-up research is one of the key features in the VliSS project and therefore discussed here briefly. We have chosen three basic topics in the methodological and empirical research:

1. Research and development of statistical methodology, to be used for evaluating the effectiveness of the web-based learning environment for different types of user groups, based on data collected from users.
2. Analysis of users' navigation strategies and their effectiveness when working with the web-based learning environment, using data generated by the web application itself and data collected in laboratory conditions.
3. Further development of the web-based learning environment, using data collected with web surveys and similar Internet-based methods.

Topic 1: To evaluate the performance of the web application, empirical data must be collected. For example, the application invites the user to respond to a built-in set of multiple-choice type questions (or problems) concerning topics examined in the current training session. This will be carried out for groups of different types of users, such as students in Statistical science attending courses on Survey sampling. Experimental designs will be planned and implemented to obtain these data. A related observational data source is provided by pop-up type web surveys targeted to the users of the application. Similar questions as in experimental designs will be asked. The data will be collected by using standard web questionnaire forms and processed by CGI procedures (Common Gateway Interface).

Each question in the web questionnaire form can be treated as a dichotomous or polytomous item. In analysing the experimental item response data, the basic Rasch model (Rasch 1960) or some of its extensions (see i.e. Gustafsson 1977; Andersen 1980; Thissen 1982; Hambleton and Swaminathan 1985; Holland 1990; Bollen 2002) can be used. Several extensions of the basic Rasch model, as well as estimation methods of the model parameters, have been proposed in the literature (Gustafsson 1977; Andersen 1980; Thissen 1982; Hambleton and Swaminathan 1985; Holland 1990). A Rasch model for partial credit scoring has been considered (Masters 1982). Kelderman and Rijkes (1994) have considered more general cases where latent variables are assumed to be multidimensional. The basic Rasch model is still widely used in educational research (e.g. Oleksandr et al. 2001). A generalized Rasch model has been proposed by Wu, Adams and Wilson (1998). In the traditional approaches, there are basically two restrictions: (1) Assumption of conditional independence of test items, and (2) Assumption of independence of respondents (corresponding to an assumption of simple random sampling of respondents from a large population).

In our environment, however, the analysis methods must be adapted to account for the various sources of correlations possibly arising from the study design. This is due to the following. A hierarchical structure can arise from the different levels prevailing in the population (for example student level and class or study group level). Accounting for this type of clustering effects is discussed for example in Lehtonen and Pahkinen (1996) and Goldstein (1995). Item clustering effect, discussed for example in Scott and Ip (2002), might have to be assumed as well if the items can not be assumed conditionally independent e.g. due to the structure of test forms to be used (for example, response to the second item might depend on response given to the first item). Ignoring these clustering effects can lead to biased estimation of model coefficients and their variances.

There are alternative methods to account for the respondent clustering effect, for example the generalized estimation equations (GEE) method introduced by Liang and Zeger (1986) and discussed in Diggle, Liang and Zeger (1994). Other methods include generalized linear mixed models discussed by McCullagh and Nelder (1989), Breslow and Clayton (1993), Wolfinger and O'Connell (1993) and McCullogh and Searle (2001).

Item clustering effect can be treated for example by viewing the problem as in modelling repeated binary measurements on subjects, and a model can be fitted using the generalized estimating equations (GEE) methodology or generalized mixed models. A hierarchical Bayes technique suggested by Scott and Ip (2002) is also an option for treating the item clustering effect. More complex item dependencies might be possible to model under graphical models framework discussed by Whittaker (1990). Such methodologies are currently under study.

Topic 2: Numerical data will be collected which include detailed information on users' navigation strategies and time spent in performing different operations in a learning session of the web-based environment. This is done by a modified version of AXS site tracking system (see Web reference 2). The response time to items in a virtual questionnaire is measured as well with CGI procedures.

There are basically two possibilities to take advantage of this data collection procedure. It is possible to model the effect of time spent in each phase of the web application on the item responses. This is in order to examine whether the time spent in each operation is a relevant explanatory variable in modelling the item response data. Another possibility is to construct the item response theory model by conditioning the item responses by the effect of time spent in an operation, and then estimate the item parameters, respondent or group parameters as well as "time" parameters.

Topic 3: This part of the study is more of a qualitative type and educational research oriented. The feedback etc. collected from users by web surveys (Dillman 2000) will be used in developing the performance of the web-based learning environment.

Testing of data collection tools has been started and some preliminary analyses have been performed.

5. Discussion

The trial version of the VliSS-Virtual laboratory in Survey Sampling has been prepared in 2002, and further development of the application is under progress. Although the number of Training Keys in the current version is limited, they cover important areas of Survey sampling. The two Training Keys introduced in this paper cover basic descriptive survey sampling and more advanced survey analysis. The other Training Keys included in the trial version treat approximative techniques for variance estimation of a nonlinear estimator (such as a combined ratio estimator) in a complex survey with the linearization and bootstrap methods. In the latter case, the main aim is to demonstrate the effect of increasing the number of generated bootstrap samples to the distribution of the bootstrap estimates. In a survey analysis Training Key, we demonstrate further the role and interpretation of interaction terms in design-based multivariate analysis with a logistic analysis of covariance model. We plan to include several additional Training Keys in the application in the future.

Research and development of the VliSS application constitutes a statistical and educational evaluation of the learning environment. The application will be developed further essentially on the basis of results obtained in this research and feedback gathered on the user needs.

REFERENCES

ANDERSEN, E.B. (1980). *Discrete Statistical Models with Social Science Applications.* Amsterdam: North-Holland.

BRANDON, B. (1996). *Computer Trainer's Personal Trainer's Guide.* Indianapolis: Que Education & Training.

BRANSFORD, J.D., GOIN, L.I., HASSELBRING, T.S. KINZER, C.Z., SHERWOOD, R.D., and WILLIAMS, S.M. (1988). Learning with technology: Theoretical and empirical perspectives. *Peabody Journal of Education,* 64, 5-26.

BRESLOW, N.R. and CLAYTON, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association,* 88, 9-25.

BROWN, H. and PRESCOTT, R. (2001). *Applied Mixed Models in Medicine.* Chichester: John Wiley & Sons.

DEWEY, J. (1951). *Experience and Education.* New York: Dover.

DIGGLE, P.J., LIANG, K-E. and ZEGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

DILLMAN, D.A. (2002). *Mail and Internet Surveys: The Tailored Design Method.2nd edition.* New York: John Wiley & Sons.

DODGE, B. J. (1995). WebQuests: A technique for Internet-based learning. *The Distance Educator*, 1, 10-13.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models. 2nd edition.* New York: Oxford University Press.

GUSTAFSSON, J.E. (1977). The Rasch Model for Dichotomous Items: Theory, Applications and a Computer Program. Reports from the institute of Education, University of Göteborg.

HAMBLETON, R.K. and SWAMINATHAN, H. (1985). *Item Response Theory. Principles and Applications.* Boston: Kluwer-Nijhoff Publishing.

HOLLAND, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrica*, 55, 149-176.

ISOMÄKI, H. (2002). The prevailing conceptions of human being in informations systems development: systems designers' reflections. Tampere: University of Tampere. (Dissertation).

JONASSEN, D. H., WILSON, B. G., WANG, S., and GRABINGER, R.S. (1993). Constructivist uses of expert systems to support learning. *Journal of Computer-Based Instruction*, 20, 86-94.

KELDERMAN, H and RIJKES, C.P. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrica*, 59, 149-176.

LEHTONEN R. and PAHKINEN E. (1996). *Practical Methods for Design and Analysis of Complex Surveys. Revised Edition*. Chichester: John Wiley & Sons.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

MASTERS, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrica*, 47, 149-174.

MCCULLAGH, P. and NELDER, J.A. (1989). *Generalized Linear Models, 2nd edition*, New York: Chapman and Hall.

MCCULLOCH, C., and SEARLE, S. (2001). *Generalized,Linear and Mixed Models*. New York: John Wiley & Sons.

OLEKSANDR S. CHERNYSHENKO, O.S., STARK, S., CHAN, K-Y., DRASGOW, F. and WILLIAMS, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36, 523-562.

RASCH, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Denmarks Pedagogiska Institut.

SCOTT, S.L. and Ip, E.H. (2002). Empirical Bayes and item-clustering effects in a latent variable hierarchical model: A case study from the National Assessment of Educational Progress. *Journal of the American Statistical Association,* 97, 409-419.

WEBSTER, F. (1995). *Theories of the Information Society*. New York: Routledge.

WHITTAKER, J. (1990). *Graphical models in applied multivariate statistics*.Chichester: Wiley.

WU, M.L., ADAMS, R.J. and WILSON, M.R. (1998). Conquest. Generalized Item Response Modelling Software. ACER-Australian Council for Educational Research.

Web references

Web reference 1: The VliSS application. http://www.stat.jyu.fi/mpss/VLISS

Web reference 2: About WebQuests – A document outlining the definition. http://edweb.sdsu.edu/courses/edtec596/About_webquests.html

Web reference 3: Java Home Page. http://java.sun.com/

Web reference 4: VRML – Virtual Reality Modelling Language – An extension of the web that allows users to manipulate 3-dimensional graphics on web pages. http://www.vrml.org/

Web reference 5: ChibaMOO Papers. Collection of papers on Multidimensional Object Oriented environments provides some glimpses of how to create virtual spaces for collaborative learning. http://sensemedia.net/papers

Web reference 6: The Applications camp.
http://edweb.sdsu.edu/edfirst/appcamp/precamp1.html

Web reference 7: Crafting Applications.
http://edweb.sdsu.edu/edfirst/appcamp/web/web.html

Web reference 8: Thinking Through Linking.
http://edweb.sdsu.edu/edfirst/appcamp/web/thinking.html

Web reference 9: Web based learning resources library.

http://www.outreach.utk.edu/weblearning/

Web reference 10: Advanced educational uses of the World Wide Web.
http://www.igd.fhg.de/archive/1995_www95/papers/89/paper.html

Web reference 11: Distance Learning Resource Network - Web-Based Instruction Examples.
http://www.wested.org/tie/dlrn/examples.html

Web reference 12: The SAS web site.
http://www.sas.com

Web reference 13: The SUDAAN web site.
http://www.rti.org/sudaan

Web reference 14: AXS-site tracking system
http://www.xav.com/scripts/axs/

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 1013–1035

TRADITIONAL AND NEW TECHNIQUES FOR IMPUTATION

Seppo Laaksonen[1]

ABSTRACT

The paper deals with imputation techniques and strategies. Usually, imputation commences truly after the first data editing, but many preceding operations are needed before that. In this editing step, the missing or deficient items are to be recognised and coded, and then decided which of these, if any, should be substituted by imputing. There are a number of imputation methods and variants of them. This paper first gives an overview of such techniques using a somewhat new approach, then considers certain newer methods in more detail, including use of neural nets. Some of these developments are tested using the two different real data sets, and the results compared with each other. Further research is needed.

***Key words*:** AID Analysis, Classification and Regression Tree, Imputation Model, Logistic Regression, Model-donor imputation, Neural Net, Non-Response, Regression based nearest neighbour (RBNN), Real-donor imputation, Self-Organising Maps.

1. Introduction

Imputation is a family of techniques for replacing missing values with values meeting one of these requirements: (i) the imputed values are expected to be close to the true values, or (ii) the distribution of the imputed values is close to the distribution of the true values, or (iii) the aggregate estimates based on these imputed values are expected to be close to the aggregate estimates based on the true values. The first requirement is the strongest. If it is successfully met, the relationships between different variables will also be close to the true relationships. It is however often difficult to meet the first requirement. This leads

[1] University of Tampere, University of Helsinki, Statistics Finland.
Email: Seppo.Laaksonen@Stat.Fi

to attempts to meet the second or the third requirement instead, which are less demanding but do not guarantee correct inference on relationships.

An awkwardness of imputations is that we do not know the true values at the time of imputing. Sometimes, such values are available later, and consequently, the success of imputations may be analysed. There is usually, however, some information, at least at the distribution or aggregate level, which can give some ideas of the success of imputations. It is also useful to benchmark the estimates of the completed data set against the estimates of the data of available cases. If the missingness mechanism could be ignorable, the estimates should be rather close to each other. However, as already pointed out, we do not usually know this mechanism well.

The replacement of incomplete values may be done either singly or multiply. The former is called *single imputation* and the latter *multiple imputation*, respectively. This paper does not deal with the latter methodology, although it could be practicable in some considered situations (see e.g. Rubin 1987, Little and Rubin 1987 or its new edition from 2002, Rubin et al 1996).

Imputation techniques may be used both for item nonresponse and for unit nonresponse. A need to use imputation for second-stage unit nonresponse arises in cases where responses are missing for some second-stage units (e.g. one or more but not all household members respond), while responses for the corresponding first-stage units (households) are available. Secondly, there are situations when some data are missing because they have not been collected in order to save the survey costs, for example. The third typical case is that of erroneous values, rejected at the editing step. Moreover, it sometimes happens that the value of a variable is known only roughly, for example, within which interval it lies. If a single value is needed, it has to be imputed within the interval (e.g. Heeringa et al 1997).

Imputation techniques have probably been used for as long as statistical data have been collected. Often, the technique has been very a simple such one as '*last carried value forward*' (a method for *cold decking*, see e.g. Kalton and Kasprzyk 1986) or average of the available values (*mean imputation*). Obviously, good guesses have also been used, leading to more or less subjective imputed values. Naturally, there is need to make such replacements objectively, and so that the solutions are documented. Moreover, more advanced methods have been increasingly exploited and tested in recent years.

Many aspects of this paper have emerged from international research projects funded by the European Union (EU). The Euredit project is one of these (see http://www.cs.york.ac.uk/euredit/). It compares the so-called new and traditional techniques for editing and imputation. The new techniques here mainly mean neural nets. In this paper we present some ideas of using self-organising maps (see e.g. Koikkalainen 1995 and 1999, Hakkinen 2001). We also consider tree-based methods which have been developed for imputation purposes by another EU project, entitled AutImp (Piela and Laaksonen 2001, Chambers et al

2001). Furthermore, we discuss the more traditional methods although these have also been progressing during recent years. Hence it is not clear which methods are new, and which old, or traditional.

The empirical results obtained with the various methods mainly take advantage of the Euredit data sets for which an independent body did the missingness but we are able to check how well we have succeeded after our imputation This gives quite good opportunities for comparing the different methods, but it is still problematic to do all comparing completely objectively, since an imputation operator may succeed better with a method he/she knows better.

Besides the different comparable and pedagogic empirical exercises, the paper aims at describing the imputation process better than is usually done. The imputation methods themselves have also been described somewhat differently than in standard books or papers (Kalton and Kasprzyk 1986, Little and Rubin 1987, Sarndal et al 1992, Schulte Nordholt 1998, Solas 2001, Marker et al 2002). The author argues that his approach is clearer than the traditional ones. Some similar ideas have earlier been presented in Laaksonen 2000.

The paper is organised so that in Section 2 we explain the role of imputations in the whole survey process. The next section concentrates on the key aspects of imputation techniques, thus revealing the approach of the author to imputations. Section 4 presents some comparable empirical results, and Section 5 concludes the paper.

2. Survey process and imputation

From the methodological perspective the survey process could be summarised as follows:

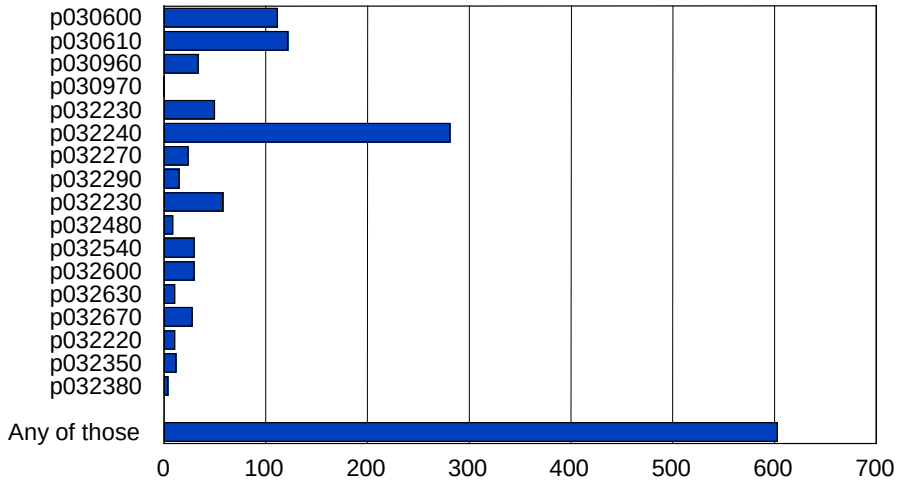
i. Users' Needs (e.g. what are the key outputs influencing imputation requirements)
ii. Survey Design including sampling design, measurement design and estimation design at the general level (which requirements for imputation need to be taken into account, avoiding such questions/variables that obviously lead to high missing and erroneous values)
iii. Exact Sampling Design
iv. Initial Weighting (design Weights)
v. Data Collection (incl. auxiliary data for imputation)
vi. Editing and Imputation
vii. Final Weighting (weights based on ignorable missingness, post-stratification, response propensity modelling for individual respondents, outlier weighting, calibration)
viii. Output Data: aggregated macro data and micro data for users (micro data should be flagged if imputed, for example)
ix. Dissemination (incl. quality information taking into account imputations).

This list shows that the imputation process has to be kept in mind all the time during the designing and processing of a survey. However, there is usually a special step for editing and imputation. In this paper, we do pay special attention to this step, although imputation may be important in other stages as well. In practice, the steps above cannot be followed straightforward. For example, although editing and imputation start close connection to data collection, the step cannot be finalised before re-weighting has been done.

Generally, imputation and re-weighting are often alternatives to each other and may often lead to practically the same result. However, there are several situations where imputation is superior to weighting. It is more flexible, no many weights are needed if the missingness mechanism is complex. A problem of re-weighting is that it only operates with observed values, but some imputation methods may obtain values outside this range.

Figure 1 gives an example where the missingness varies from one variable to the next. Although the missingness rate is low for each particular variable, it is much higher in its multivariate meaning. If we were here to only use weighting and units with full information, we would lose a much higher number of observations. This leads us to look for a reasonably good imputation method (for an analogous case, see Narhi, Laaksonen et al 2001).

Figure 1. Number of missing values for a group of income variables in the Finnish European Community Household Panel 1996



3. Imputation Process

We consider here that the imputation process consists of the following 6 steps: (i) *Data editing process*, (ii) *Supply of auxiliary information*, (iii) *Building of a good imputation model*, (iv) *Imputation task* (may lead to new editing), (v) *Estimation including point estimates, sampling variance and imputation variance,* (iv) *Outputs* from imputations including flagging, and dissemination to aggregate statistics and micro data users.

This paper concentrates on steps (iii) and (iv), and the focus is on the minimising of bias. However, we know very well that there are still remaining problems in variance estimates, such as estimating the imputation variance in complex situations. In most published papers successful variance estimates have only been developed for some imputation methods (see e.g. Rao and Shao 1992, Rubin et al 1996, Shao 1997, Shao 2002, Lee et al 2002).

We are not here considering data editing, which usually precedes the imputation process but should be well integrated with it. The pre-editing process identifies the values, that need to be imputed. It is possible that a new editing, *post-editing*, is needed later on after imputation, since the imputed values do not necessarily fulfil the edit rules unless the imputation task has been done conditional to them which may be difficult. It should be noted that *pre-imputation* is an essential part of editing, especially if the desire is to use selective or significance editing (Lawrence and McKenzie 2000). In this paper, we do not pay attention to pre-imputation, however.

Before proceeding to look at the two most important steps in imputation, it is necessary to remember that there is need to provide as versatile, good, and up-to-date auxiliary information as possible, preferably at the individual unit level, even though aggregated data could also be exploited (see the classification of auxiliary data in Laaksonen 1999, and its revised version Laaksonen 2002). This work will have to continue during subsequent tasks if reasonable results have not been achieved with the available variables and their initial forms.

Next, we go on to consider the most crucial aspects of the *imputation model,* on the one hand, and the *imputation task*, on the other. Although we present them separately, these two steps should be integrated, and interactively built. Experience has shown that when looking at the first imputed values against some available benchmarking information, the results are not necessarily satisfactory. Hence, for example, a new specification of a model may be tried, and maybe better results achieved, consequently. Finally, a best combination of the whole operation – that is, the model and the imputation task – must have be chosen for this particular application.

3.1. Imputation model

The dependent variable of the imputation model may be of the two types:

i. Variable being imputed, or
ii. Indicator of the missingness mechanism of the variable being imputed.

If y is this variable, then in the case of (i) the model is constructed using the non-missing part of the data set (say for r units), but the explanatory variables of the model should be available for the missing part as well (say for n units). The initial variable may be transformed by, among other things, logarithm, and it may be also categorised. A practicable strategy for categorisation in cases where non-negative variable y has many zero values is to binarise it so that if $y>0$ then the new variable $z = 1$, whereas in other cases $z = 0$. This occurs fairly often in business surveys, in particular (see e.g. Laaksonen 2000). This leads to the exploitation of two models. After the above mentioned first step, a new model is to be built for those with non-zero values, these being either real or imputed ones.

Respectively, in the case of (ii), the model is to be built for n units and the response variable may, for example, be:
= 1 if the value is non-missing, and
= 0 if the value is missing.

As in the case of (ii) the explanatory variables are ones that are available for the full data set (n units).

The imputation model may be specified in different ways. It is impossible to give a completed list of these specifications. In some cases, the model behind imputations is so simple that one can not see any model. For logical imputations, a statistician may use a known function, or an edit rules specification, for example. In some sense, the model may resemble a good guess, derived from the experience of this survey process.

When advancing to more complex techniques, the model needs an estimation using the available data. A typical specification is a linear statistical (regression) model. In very simple cases, such a model may only include a noise term, or a constant and a noise term or one explanatory variable with or without a noise term. In these cases, only little auxiliary data are available. As the amount of auxiliary data increases, more complex and multivariate models should be used. Logistic regression is often used for binary response variables. In principle, many types of traditional or untraditional statistical models may be tried as imputation models.

Naturally, new modelling techniques may be exploited, such as classification tree for categorical variables and regression tree for continuous ones, respectively. Recently, some efforts have been made to test neural nets for editing and imputation under the Euredit project. Over a number of years the Jyvaskyla University group has been developing special software, called *Neural Data Analysis* (NDA), which specialises in the *self-organising maps* methodology (*SOM*). Currently, this NDA group together with Statistics Finland methodologists has been extending the existing technique for editing and imputation (later referred to *NDAEI*). Some first NDAEI experiences are presented in this paper. Besides the SOM, the Euredit project is also testing the three other neural nets technologies, that is, Multilayer Perceptron (MLP),

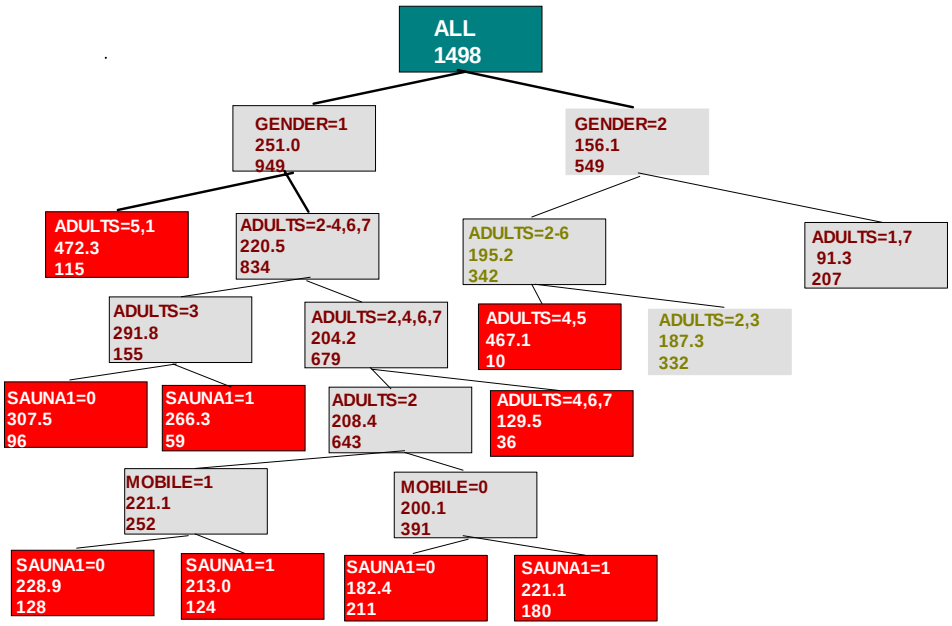
Correlation Matrix Memory (CMM, AURA) and Support Vector Machine (SVM), see Euredit website (http://www.cs.york.ac.uk/euredit/). .

There are special features both for constructing an imputation model and for imputation itself, consequently. These models may be built either, (a) for the whole data set considered, or (b) independently for certain sub-sets of the data set.

Imputation cells

The latter sub-groups are often called *imputation cells*, or *imputation classes*. The nature of such cells is also similar to adjustment cells or response homogeneity groups of reweighting techniques. In the case of (a), this kind of imputation cell may be an explanatory variable. Ideally, an imputation cell is homogenous or the missingness mechanism is ignorable or at least MAR (see e.g. Rubin 1987) within this cell.

Figure 2. The model is a regression tree, *consumption of alcohol* of Finnish households is the target variable and 4 Explanatory Variables are tried. The right-hand tree (gender=2) is not shown in full.



A specific part of the imputation process is to define these imputation cells as correctly as possible, but how to create them, it is a major question. In an ad hoc approach, an imputation operator just makes his/her choice by intuition and experience of parallel situations. A more advanced way is to exploit any good modelling technique for this purpose. For example, logistic regression may be used so that the dependent variable is = 0 if the value is missing and = 1 if it is not

missing. The estimated propensity scores are next divided into quintiles and these are used as the imputation cells (this is a solution in Solas). Naturally, many other techniques for finding such cells may be used under traditional models. Certain new methods, such as classical tree-based ones and SOM may be used to automatically provide the possible imputation cells for further use. However, there are no automatic criteria for finding the optimal combination of cells. In Figure 2 we give an example that is based on regression tree.

We do not present the regression tree methodology in detail here (see e.g. Breiman et al 1984, Piela and Laaksonen 2001 or Chambers et al 2001 and software WAID, available from Statistics Netherlands). The methodology itself is some decades old, it is based on a computer-intensive algorithm which first splits the target population into two independent sub-groups or nodes. Next, each of these nodes is split again into two nodes, and in the end there will so-called *terminal nodes*, which are used as imputation cells. It is possible to give several criteria for the splitting and for the extent to which is should be taken. For example, the maximum number of terminal nodes, or the minimum number of observations for such a node may be given. It is also possible to weight the algorithm, for example so that the outliers have a smaller weight; this means a robust method.

In Figure 2, there are 8 terminal nodes for males (gender =1), and one node for females since this part is not shown in full. Some terminal nodes are found at quite an early stage (e.g. if head of household = male and if the household has 1 or 5 adults), whereas some terminal nodes are derived from all the 4 explanatory variables (e.g. head of household = male, number of adults = 2, they have no mobile phone but they do have sauna). The figure also shows the average of alcohol consumption (these are in the above particular cases = 472.3 and 221.1) and the number of observations (= 115 and 180).

The technology for the NDAEI is from the imputation point of view analogous to tree-methods, but the algorithms behind this SOM are more complex and comprehensive (Hakkinen 2001, Koikkalainen 1995 and 1999, etc.; for general background on neural nets, cf. e.g. Kung 1993 or Nordbotten 1995). A specification is called *Tree-Structured Self-Organizing Maps (TS-SOM)*. This technology starts from the full data set, called *root – one neuron*, and next creates the first sub-groups, called *SOM layer 1 – four neurons* (these correspond to clusters or imputation cells) and each of these into 4 sub-groups, called *SOM layer 2 - 16 neurons,* and so on. These layers or levels, and neurons or cells, may be examined graphically and standard statistical tables and indicators may be provided. An imputation operation may be done within each neuron. A user has to choose which level of layers to apply. It is possible to use a certain level for one part of the data set, and another level for the rest.

3.2. Imputation Task

We here use terminology that consists of the only two basic techniques for imputation conditional to the imputation model:

i. In the case of *model-donor imputation,* the imputed values are *directly* derived from a (behavioural) model.
ii. In the case of *real-donor imputation,* the imputed values are *directly* derived from a set of observed values, from a real donor respondent, but still are *indirectly* derived from a more or less exactly defined model.

In practice, the final method may include both of these alternatives. We next present some specifications for both types of imputation tasks:

Options for the model-donor method

In this case, an imputed value is either

i. a predicted value of the model (*deterministic solution*), or
ii. a predicted value of the model plus a noise term (*stochastic solution*).

It is not always clear how this runs. If an ordinary regression model has been used, then the predicted value is just an imputed value (often called *regression imputation*, and in more simple cases *mean imputation* or *ratio imputation*).

Where a predicted value is a probability, say p_k , thus within interval (0, 1), then there are two alternatives for deciding whether the imputed value is to be either 1 or 0:

(a) to create an uniformly distributed random variable within (0, 1), say u_k, and then the imputed value = 1 if $p_k > u_{k,} = 0$ otherwise,

or

(b) to use external information about the level the probability should be (e.g. less than 0.5 then the imputed value = 0, and =1 otherwise). This external information may be derived from the training set or just based on a good guess.

Case (ii) for model-donor imputation requires some assumptions on the distribution of the noise term. The noise term should be fitting well with real unknown observations (in order to achieve proper imputation).

The most usual way for continuous variables is to use *normal distribution* so that the mean and the variance of the noise term are estimated from the real data. In practice distributions are however often skewed and may have to be truncated. The correct way of doing this has not been studied much, and thus needs more research. I have tried to look at the distribution of residuals. In my experience, even one standard error or one root mean square error is enough, sometimes even too large, but it depends much on the predictability ability of the imputation model. If no bounds are used, the results may be problematic. For example, if the Solas software is applied automatically to ordinary business survey data with skewly distributed variables, it is likely that some big outliers containing non-acceptable negative values are created (Chambers et al 2001).

If an explicit model has been used, and the residuals for the observed units have been estimated, then it is useful to exploit these either by

- choosing one of the residuals randomly (model-donor) or
- choosing a residual of a nearest (near) neighbour of a unit with a missing value (this comes under real-donor methods, see Section 4 too).

On the contrary, for categorical variables, empirical or estimated distribution should preferably be used so that *the distribution of the imputed values will be equal to the observed distribution* (with random error since in this operation the uniformly distributed random variable has been used as a technical tool as above). Some 'imputeurs' prefer to use 'mode imputation,' that is, the category with the highest frequency/probability will be used as an imputed value. Note that this method is usually applied within small imputation classes, where it may prove more realistic.

Options for the real-donor method

The model is used thus to find a donor, whose' value has been used as the substitute for the missing one. In principle, the donor may be chosen by various criteria, but it is most natural to choose it from *the neighbourhood of the unit* with a missing value. If a nearest donor and the regression model without noise term have been used, or fixed distribution has been applied, this leads to deterministic imputation. In all other cases, some random elements are included in the process, that is

- a model has a random noise term, or
- an imputed value has been randomly selected from the neighbourhood (like in the case of so-called random hot decking).

What are the metrics for finding the neighbourhood is a big question with real-donor methods. Some suggested solutions include:

- Euclidean distance: How to scale the variables is an important element under this method especially if several auxiliary variables are used. Very often all variables are scaled for the same level, for example within an interval (0,1) and next each of these have been weighted equally or unequally. It will be subjective to some extent to determine the weights without testing them against a good training data set.
- Gini index of categorical variables: this is the metrics used in the classification tree of the WAID software (Chambers et al 2001).
- Edit rules must have been satisfied.
- When using certain methods, like random hot decking within imputation cells it is assumed that all distances within this cell are equal.
- Robust measures for each cases.
- *Predicted values of the estimated statistical model with or without random noise term* (Laaksonen 2000). Note that these same values may also be used as imputed values in model-donor methods. As pointed out in the context of

model-donor methods, it is not clear how to truncate this noise term. An advantage from the adding of the noise term is that it gives a good tool for multiple imputations. It seems to be good for single imputations in certain cases, too (e.g. it scatters piles and gives more opportunities for finding different neighbours). An advantage of this metrics is that the auxiliary variables will be weighted using an empirical training data set of the respondents, and this thus leads to an objective weighting solution (cf. Euclidean approach).

An end of this section I present some pedagogic examples of my approach to imputation, which essentially integrates the construction of an imputation model, and an imputation task, respectively. Let us consider a plain linear regression model

$$y = \beta_0 + \beta_{1k} x_{1k} + \beta_2 x_2 + \varepsilon, \quad \text{in which}$$

ε = random noise term, y = survey variable, x_{1k} = domain for domain k (what may be used as an imputation cell) and x_2 = auxiliary variable. The estimates for the parameters are denoted by b_0, b_1, b_2 and e, respectively.

If this model is reduced so that the estimated equation is $y = b_0$, then it is called *mean imputation*, or if the variable is of a ratio type, *ratio imputation*, respectively (median may be also possible). This does not take advantage of any auxiliary variable but it presumes that the missing units are similar to the average of the respondents.

If the model is reduced to $y = \varepsilon$, then imputation may be done by using observed residuals or theoretical 'residuals' assuming that these follow a certain distribution such as normal distribution, but the imputation may be done either using a real-donor or a model-donor technique. If just these theoretical values have been used, it is a model-donor technique, whereas if one chooses each imputed value randomly from the set of observed residuals, the technique is a mixed one, but nevertheless stochastic, since it includes elements of both approaches. There is also an option to assume that this random term is uniformly distributed and this holds true for the missing units too. If we next choose randomly one respondent (real-donor) and give this value to a randomly selected missing unit, this leads to the method which is usually known as *overall random hot decking* given that the same donor may be chosen as many times as the process allows (with replacement). It is also possible to make this selection without replacement but this leads to worsening problems while the missingness rate increases. If the rate is higher than 50 percent, all donors have already been used.

If our estimated model is $y = b_0 + b_{1k} x_{1k} + b_2 x_2$, and these values are directly used as imputed values, it is often called (pure) *regression imputation*. But if we take the nearest neighbour metrics from this estimated model, it leads to a real-donor method which Laaksonen (2000) has called *regression-based nearest neighbour hot decking* (RBNNHD), for which we here use the term RBNN

shortened by ignoring the last two words. Correspondingly, the noise term with theoretical or observed residuals may be added.

If the model $y = \beta_0 + \beta_2 x_2 + \varepsilon$ has been estimated independently for each category of variable x_1, then the respective methods could be called 'cell-based,' for example.

4. Comparative results from the two data sets

To concretise our approach to imputation methods we present in this section examples from the two quite different situations. These examples do not cover the methods extensively, since in both cases the variable being imputed is a continuous one. On the other hand, these two circumstances differ much from each other. For example, in both cases the distribution of the variable is rather skew, but much skewer in the first case than in the second one. This leads to special difficulties in imputations. In contrast, the second case is more difficult from the point of view of auxiliary variables. Using these examples we aim at giving some ideas about the crucial points of imputations. We are not able to say definitely which method is the best in each case, since the conclusion also depends on the target the users have set to the imputation. However, we see that certain methods are superior to others.

The data sets used in these examples are constructed by an anonymous person as a special task for the Euredit project. Hence we cannot know what the applied missingness mechanism principles have been. Fortunately, we can check our results afterwards against the true values.

The first data set is from t*he UK Annual Business Inquiry (ABI)*, which is based on a typical business survey design. The data are available from two successive years, however so that there are only about 25 percent of the units from each year, these being mainly big firms that are included in the survey at certainty. Typical variables of interest are *survey turnover*, *employment cost*, and *investment*. In our example, we only consider *survey turnover*, which is one of the key variables of the survey. A possible strategy would be to first impute the missing values of this variable, and next use *this completed variable* as an auxiliary variable when imputing such variables as total employment cost and total investments. There is a need to check the components of each of these variables, and this leads to further imputations using these 'total' variables as auxiliary variables, among others. This kind of strategy is called *sequential imputation*.

Although it has been seen that this key variable may have a high influence on several survey results, there are not many auxiliary variables available in this data set when imputing this variable itself. Fortunately, especially *turnover derived from a register* is available. The concepts of turnover are not very source-dependent, but more differences are derived from the different reference periods. An enterprise, for example, is not working at the calendar-year principle. The

register turnover is also often based on older information, since there may be a lag in the updating of the register. Nevertheless, the correlation between these variables is high.

The second data set is derived from *the Danish Labour Force Survey (DLFS)***.** There is only one variable with missingness in this data set, that is, *individual yearly income*. The data are fairly typical for such surveys since the data set contains several characteristics of individuals such as gender, age, employment status, living area and education level. However, there are no variables that relate well to personal income at the individual level.

We present our results in the following two sub-sections.

4.1. Imputations for TURNOVER 1998 of the 1997-1998 ABI data set

Our best auxiliary variable in all imputation models has been *register turnover* from the same year. In addition, we have included in the cross-sectional models two categorical variables, that is, *level of register employment* (6 categories) and *industry class* (3). In the case of panel models, we have also used the following auxiliary variables from the previous year: *survey turnover, register turnover, total taxes paid, total purchases*.

Table 1 gives the results of the 9 different exercises. The notation of the table is an illustration of our strategy for imputations. We thus see that at least three factors specify an imputation method: the data used for modelling, the model specification itself (these two thus cover 'imputation model'), and the way imputation has been done. Naturally, if variance estimates have been provided, this would require a new column in the table.

The background of each method has been explained in Sections 2 and 3, but the method called *Residual RBNN* requires an additional description, since it was created for this experiment. The basic idea behind this method is the same as that of the RBNN. We first estimate a multivariate regression model. Next, we search a nearest neighbour for each missing unit using the metrics of the predicted values. Finally, we do not take the real values from this donor but we take the estimated residuals from this donor instead, and add these values to the predicted values of the missing unit. This methodology must be examined further, but logically, it seems to be advantageous, at least since it is more robust against influence from outliers, correlated residuals, and similar deviations of reality from the model.

The models for TURNOVER seem well fitting in all cases. Panel models were only possible for such units that were included in the sample in both successive years. In these models, when using linear scaling, the R-squares were around 95 percent, whereas when using log-scaling the respective fits were around 90 percent. The fits for cross-sectional models were somewhat lower, around 80 percent. High R-square values are, however, not really surprising for untransformed, size-related variables with a large natural variability. Goodness of fit is a popular indicator for the success of imputations, and here it apparently

works at least in the sense that the results when using panel data are, on average, better than when only using cross-sectional data. This is not always guaranteed, but in this case the panel businesses – mainly big ones – seem to differ from the cross-sectional ones, and hence the same imputation model is not advantageous to both.

Table 1. The ABI 1997-1998 test results for variable TURNOVER in the imputed part of the sample.

RTMSE = root means square error of regression model. One big true value – outlier – has been excluded in the figures in parenthesis. The number of imputed values = 103 (102), and the whole sample size = 5594.

Data for model	**Model type**	**Imputation task**	**Quality indicators**			
			Mean	**CV**	**Maxi-mum**	**Mean absolute error**
True values			**76274 (25620)**	**686 (385)**	**5242956 (887613)**	**0 0**
Cross-section	Linear	Predicted	148067 (27417)	829 (322)	12454429 (754133)	77236 (7293)
Cross-section	Log-linear	Predicted	44822 (20895)	563 (334)	2485495 (606813	34729 (8036)
Cross-section	Log-linear	Predicted plus 1 RTMSE	66257 (43330)	470 (481)	2404862 (1940782)	59919 (32663)
Cross-section	Log-linear	Predicted plus 1.5 RTMSE's	155100 (100425)	650 (689)	9944179 (6928171)	150528 (90468)
Cross-section	Linear	Predicted plus residual RBNN	125698 (28194)	793 (415)	10071016 (1053046)	54488 (8490)
Cross-section	Log-linear	Predicted plus residual RBNN	103626 (30497)	725 (386)	7562812 (1051774)	34670 (8715)
Panel	Linear	Predicted	98123 (31426)	701 (399)	6901249 (1155652)	24161 (8141)
Panel	Log-linear plus linear	Predicted	80091 (28455)	666 (360)	5347064 (903708)	6032 (5071)
Panel	Log-linear	Predicted	78311 (26415)	681 (384)	5347064 (903708)	5565 (4600)

The log-scaling seems to give, on average, better results than linear scaling does. Note that the R-squares are lower for log-turnover but as well known, a high R-square does not automatically lead to the best results with regard to predictability. In general, it is probably not too unusual for log-transformation of

size-related business variables to yield a better performance but a smaller R-square, as is the case here. The best overall results are achieved when the imputation model is based on panel data and log-turnover, and when next model-donor imputation is applied.

The results for the panel case are very satisfactory, but we wanted to compare some techniques for the cross-sectional case, since this is more difficult, and it is often in practice the only possible method. Now, there are far more differences in the results. When comparing the results with and without an outlier, the clear conclusion is that the success depends much on the success with this outlier. Some methods are very un-robust, especially the technique which tries to take advantage of the distribution of 'theoretical residuals.' Even when we assume that the random noise term is normally distributed with zero mean and the bound for the root mean square error is equal to 1.5, some imputed values may be very big. This naturally depends on the random number chosen, but of course we cannot subjectively choose a "random" number so as to meet one of our objectives.

It is interesting that the new 'residual RBNN' method succeeds quite well, its specification for the log-scaled model is the best cross-sectional result. This is largely due to its better robustness against influence from outliers as compared to model-donor techniques. However, this method is not equally good when one big value is excluded. Note that the exclusion of one value after all imputation procedures is not fully fair to the methods since all the models have been done without taking this information into account.

As expected, the variation coefficients (CVs) are underestimated for model-donor techniques without noise term. This also is concerned the panel data results with an outlier. There is one strange exception, the 'cross-section-linear-predicted' technique which gives too large an estimate for the whole data set, but too small an estimate for the 'non-outlier data set'. This example shows how un-robust such a technique may be.

It is good to use graphics for checking the success of imputations. In practice, it is not possible to know the true values as in our case, but some analogous benchmarking numbers are usually available. Figures 3 and 4 illustrate the two opposite situations from Table 1, the first in Figure 3 is the worst and the second is the best, respectively. We can easily see from the figures that the second is better, but the first imputed values do nevertheless still fit quite well with the true values.

Figure 3. True values (x axis) against imputed values (y axis) for an inadequate technique, cross-section – linear – predicted (see Table 1)

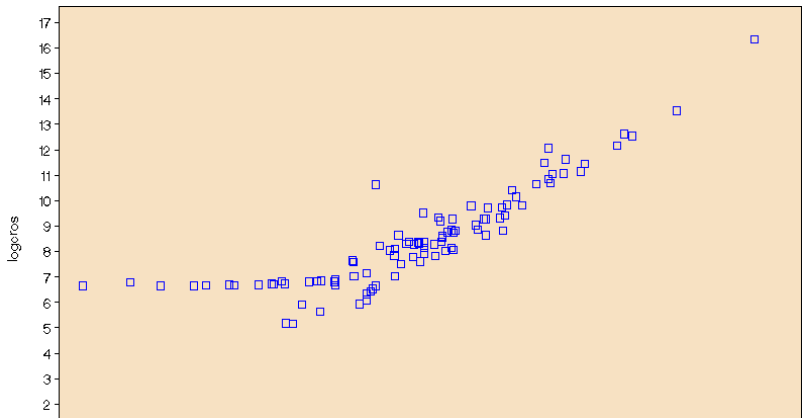
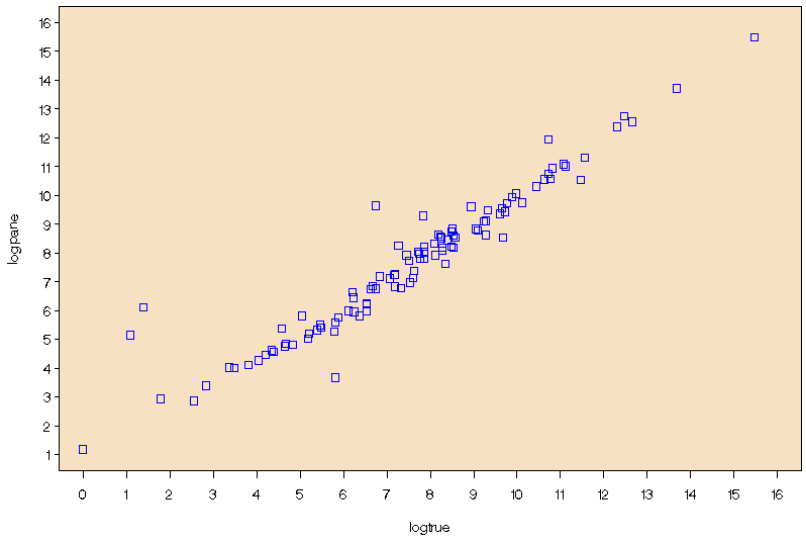


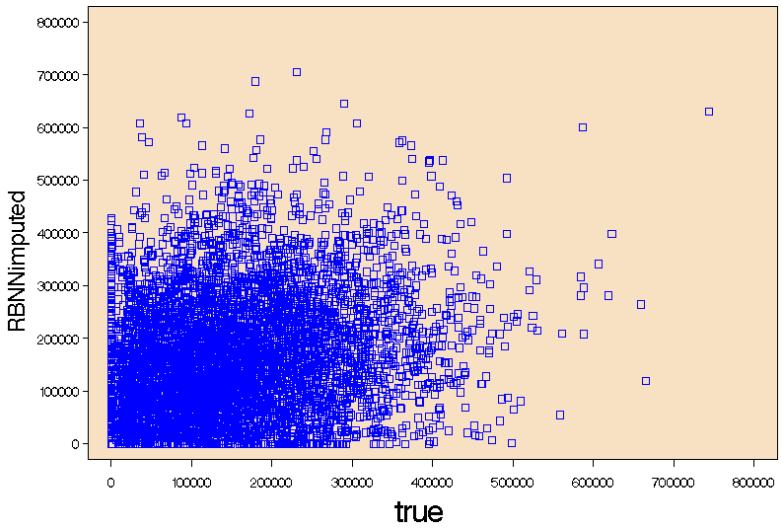
Figure 4. True values (x axis) against imputed values (y axis) for an adequate technique, panel – log-linear – predicted (see Table 1)



4.2. Imputation for INCOME in the Danish Labour Force Survey (DLFS)

It is good to start by looking at the analogous scatter plot into Figures 3 and 4, in the case of the DLFS, in Figure 5. This is based on a RBNN method, but many other respective scatter plots look similar in this rather big data set consisting of 200000 persons. The value of variable INCOME is missing for 53677 persons (missingness rate=23.8%).

Figure 5. RBNN imputed values against true values in the Danish Labour Force Survey for a 10% random sample of the missing values



We compare the results from the four types of methods with each other in Table 2: (i) regression or MLP types of model-donor techniques, (ii) SOM-based method using NDAEI software, (iii) RBNN method, (iv) random hot decking or real-donor method with random draw.

First, we want to note that the last method is done for benchmarking in the sense that it is expected that all good methods will give better results than this. Nevertheless, this is a rather demanding method for benchmarking because it should give quite similar results as derived from the available units. If the missingness mechanism is of the type *missing completely at random (MCAR)* this method should succeed well. Thus, if any good imputation model cannot be built, the random hot decking should be as good as the other methods.

The R-squares for linear income are as in the case of the ABI, higher than those for log-linear income, around 35-40 per cent in the former and around 10 per cent in the latter. We cannot explicitly tell how well the SOM models[1] are fitting, but we have looked at several specifications in these cases, too, and some of these have been included in Table 2.

[1] The results for SOM have been based on the work in the Euredit project. P. Piela from Statistics Finland has provided the results of Table 2 for this paper.

Table 2 includes several indicators of the quality of imputation, since one indicator is not enough, for example mean absolute error which seems to work best for direct model-donor methods. This is due to the fact that no model does fit well, and hence the average types of model values, which are somewhere in the middle, are not very far from any true value, but almost all are quite far. On the contrary, such imputation techniques do not give any correct distributional figures (CVs, quantiles) which are extremely important for the users of this type of a survey. Hence we cannot recommend model-donor methods although they are good if based on one criterion.

When taking into account distributional requirements the SOM-based methods and the RBNN methods, respectively, are the best. There are no big differences between each other. However, if the number of neurons, and the layer level are high, the results are not automatically improving, and maybe even worsening. This is due to the problem that the imputation cells will become smaller and smaller and the possibilities for finding good near real donors will worsen. This was also observed when using standard tree-based methodology (Piela and Laaksonen 2001). The optimum number of layers, or terminal nodes, is not very high, but high enough. It is not easy to find an objective stopping criterion for the tree building.

It is finally interesting that the RBNN method is somewhat better than the SOM method when using the criterion "absolute relative error." Obviously, this technique is slightly better for exploiting poor auxiliary data. The reason for these results should be investigated further.

5. Conclusions

In this paper, we consider imputation within a broader framework than is ordinarily done. A key point is that when speaking about imputation we need to pay much attention to the model behind the imputation task. Moreover, it is often difficult to well estimate the parameters of this model so as to succeed in the imputation. The basic target in the estimation of a model is to achieve good predictability. A specific problem is that it is difficult to estimate a model over the whole range of true observations, especially for typical business survey variables with skew distribution. This has been demonstrated in the first example of the paper.

The example also shows that if a model is well fitting, it is possible to succeed with model-donor methods so that the predicted values of the model are used as imputed values. On the other hand, if a model is not so good, real-donor techniques are often superior, or at least, show their advantage in being more robust than model-donor techniques. Missing observations for true outliers are, however, problematic under all circumstances, and it is necessary to try to collect data for them rather than to impute.

The second example in the paper is simpler in the sense that there are no hard outliers. On the other hand, the imputation models are much worse. The data set is very big which helps in finding reasonably good real donors. For these two reasons real-donor techniques succeed much better than model-donor ones do. We do not exclude the possibility of finding a good model-donor technique for such a case. It requires excellent exploitation of information on the estimated random noise term of the model. If we could know its behaviour completely, for example, whether it is normally distributed with a known mean and variance, we would be certain to succeed, but this does not occur often in real life.

As a conclusion, it is awkward that there are always some uncertainties or black boxes in data that have been completed by imputation. This is sometimes a reason even the reason for avoiding these techniques, but this is not necessarily a good solution, since it may impair the quality of the data. The criticism against imputation is largely derived from the fact that it is not always successful at the individual value level. This is demonstrated well by our second example. Fortunately, our best results in this case are very good at the distributional level. It should be noted that income distribution is precisely of the highest interest for users, and hence we may be happy with our results. On the contrary, it would be correct not to recommend to the use of imputed data for an individual level analysis in this case. Instead, the best results of the first example may be used in such analyses, too.

The truth is that imputation has been used implicitly in the survey process, especially as a part of data entry and data editing. The editors may have logically deduced which values are the most correct ones when observing small errors and holes. Thus, good guess techniques have been used for imputation. The principles of these techniques are rarely well documented and it is difficult to check afterwards how correct these operations have been. It is much better to use explicit imputation methods. Imputation could be thus used explicitly and also for quality reasons much more than is done today. There is great need to further develop imputation techniques in the future. Good auxiliary data are of high value for imputations, as well as for data editing, weighting and estimation. Hence these kinds of data services should be improved for the survey process (see Laaksonen 2002).

Finally, I try to return back to the title of this paper which promises discussion on traditional and new techniques. It is not very clear what the traditional and new techniques are, although such methods as mean imputation, ratio imputation, last carried value forward and simple random hot decking may be classified as traditional ones. In some sense, the term ‘new techniques’ is used when the model applied in the imputations is considered to be new in the sense that it has not been traditionally used for this purpose. For example, the purpose of the Euredit project is to develop and evaluate new methods for editing and

imputation, and in this context it mainly means that neural nets are especially experimented with in developing imputation modelling.

We may, however, interpret these new methods more broadly. An idea behind these techniques could be higher level of automation. In this sense, it is easy to include tree-based methods in the family of new methods. On the other hand, many traditional models, for example, multivariate regression models, have not been well exploited in imputation uo to now. Hence I include here RBNN types of applications in new imputation methods. This kind of classification is not, of course, the most important. It is much more important is to further develop new techniques for imputation, and especially the tools for implementing them into practice.

REFERENCES

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, C.A.

CHAMBERS, R.L., HOOGLAND, J., LAAKSONEN, S., MESA, D.M., PANNEKOEK, J., PIELA, P., TSAI, P. and de WAAL, T. (2001). The AUTIMP-project: Evaluation of Imputation Software. Research Paper 0122. Statistics Netherlands.

HEERINGA, S.G., LITTLE, R.J. and RAGHUNATAN, T.E. (1997) Bayesian estimation and inference for multivariate coarsened data on U.S. household income and wealth. Invited Paper *for the 51st Session of the ISI*, Istanbul.

HORTON, N.J. and LIPSITZ, S.R. (2001) Statistical Computing Software Reviews. The American Statistician*,* 55, pp. 244-254.

HAKKINEN, E. (2001). *Design, Implementation and Evaluation of the Neural Data Analysis Environment*. PhD thesis. Jyvaskyla University Library, Jyvaskyla, Finland.

KALTON, G. and KASPRZYK, D. (1986) The Treatment of Missing Survey Data. Survey Methodology 12, pp. 1-16.

KOIKKALAINEN, P. (1995). Fast Deterministic Self-Organizing Maps. In Fogelman-Soulié, F. and Gallinari, P., eds., *Proc. ICANN'95, Int. Conf. on Artificial Neural Networks*, volume II, pp. 63-68, Nanterre, France. EC2.

KOIKKALAINEN, P. (1999). Tree Structured Self-Organizing Maps. In Oja, E. and Kaski, S., eds., *Kohonen Maps*, pp. 121-130. Elsevier, The Netherlands.

KUNG, S.Y. (1993). *Digital Neural Networks.* Prentice Hall, Englewood Cliffs, NI.

LAAKSONEN, S. (1991). Adjustments for Non-response in Two-year Panel Data. *The Statistician.* Great Britain 40, pp. 153-168.

LAAKSONEN, S. (1999) Weighting and Auxiliary Variables in Sample Surveys. In: G. Brossier and A.M. Dussaix (eds). "Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches," pp. 168-180. Dunod. Paris.

LAAKSONEN, S. (2000). Regression-Based Nearest Neighbour Hot Decking. *Computational Statistics* 15, 1, 65-71.

LAAKSONEN, S. (2002). Need for High Level Auxiliary Data Service for Improving the Quality of Editing and Imputation. *Paper for the UNECE Work Session on Data Editing in Helsinki, 27-29 May.* Available. E.g., on the UNECE website: www.unece.org.

LAWRENCE, D. and McKENZIE, R. (2000) The General Application of Significance Editing. *Journal of Official Statistics* 16, pp. 243-254.

LEE. H., RANCOURT, E. and SARNDAL, C-E. (2002). Variance Estimation from Survey data Under Single Imputation. In: *Survey Nonresponse* (eds. R. Groves, D. Dillman, J. Eltinge and R. Little). Wiley Series in Probability and Statistics. pp.315-328.

LITTLE, R. (1988) Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics* 6, pp. 287-297.

LITTLE, R. and RUBIN, D. (1987, 2002). *Statistical Analysis with Missing Data.* John Wiley & Sons.

MARKER, D.A., JUDKINS, D.R. and WINGLEE, M. (2002). Large-Scale Imputation for Complex Surveys. In: *Survey Nonresponse* (eds. R. Groves, D. Dillman, J. Eltinge and R. Little). Wiley Series in Probability and Statistics. pp.329-341.

NORDBOTTEN, S. (1995). Editing Statistical Records by Neural Networks. *Journal of Official Statistics* 11, 391-411.

NARHI, V., LAAKSONEN, S., HIETALA, R., AHONEN, T. and LYYTINEN, H. (2001). Treating Missing Data in a Clinical Neuropsychological Dataset-Data Imputation. T*he Clinical Neuropsychologist, 380-392.*

PIELA, P. (2001). Introduction to Self-Organizing Map Modelling for Imputation – Techniques & Technology. Contributed paper for the ETK/NTTS Conference, organised by Eurostat and The Joint Research Center. June, Crete, Greece.

PIELA, P. and LAAKSONEN, S. (2001). Automatic Interaction Detection for Imputation – Tests with the WAID Software Package. Conference Proceedings of the Federal Committee on Statistical Methodology, Washington, November.

RAO, J.N.K. and SHAO, J. (1992) Jacknife Variance Estimation with Survey Data Under Hot Deck Imputation. Biometrik*a* 79, pp. 811-822

RUBIN, D. (1987*)* Multiple Imputation in Surveys. John Wiley & Sons.

RUBIN, D. and the papers and the discussion by B. Fay, J. Rao, D. Binder, J. Eltinge and D.

JUDKINS (1996) Multiple imputation after 18+ years. Journal of the American Statistical Association 91, pp. 473-520.

SARNDAL, C-E., SWENSSON, B. and WRETMAN, J. (1992) Model Assisted Survey Sampling. Springer.

SARNDAL, C-E. (1996) For a better understanding imputation. In: S. Laaksonen (ed.). International Perspectives on Non-response. *Statistics Finland Research Reports 219*. pp. 7-22.

SCHAFER, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall.

SCHULTE NORDHOLT, E. (1998) Imputation: Methods, Simulation, Experiments and Practical Examples. *International Statistical Review*, 66, pp. 157-180.

SHAO, J. (1997) Variance Estimation for Imputed Survey Data With Non-Negligible Sampling Fractions. Invited Paper for the *51st Session of the ISI, Istanbul*.

SHAO, J. (2002). Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. In: *Survey Nonresponse* (eds. R. Groves, D. Dillman, J. Eltinge and R. Little). Wiley Series in Probability and Statistics. pp.303-314.

SOLAS (2001). *Solas for Missing Data Analysis 3.0.* Statistical Solutions, Ltd. Cork, Ireland.

Table 2. Results of the DLFS for INCOME in the imputed part of the sample

Imputation		Quality Indicators							
Model	**Task**	**Mean**	**CV %**	**95%**	**75%**	**MD**	**25%**	**MAE**	**MAXE**
True values		**158108**	**67.8**	**362639**	**221423**	**140971**	**76691**	**0**	**0**
Linear Regression	Predicted	170287	36.1	272202	215989	168978	122344	73678	520718
SOM, 4th Layer	Centroid	169177	33.6	259743	204815	173937	122951	73079	564345
MLP	Model-donor	166806	41.3	277061	219764	162920	109096	67279	518621
SOM, 0th Layer	NN	158068	68.2	363148	220904	140105	77004	90865	664950
SOM, 1st Layer	NN	158094	68.2	362674	221534	139715	76869	91072	664950
SOM, 2nd Layer	NN	159799	67.8	363374	224541	141657	77914	91593	671202
SOM, 3rd Layer	NN	159408	67.9	363712	223804	141616	77779	91639	664950
Linear, no random term	RBNN	159441	67.5	360779	223921	142718	77601	87592	672713
Linear, random term	RBNN	159448	68.0	363389	224346	142403	76785	88208	793438
Log-linear, no random term	RBNN	160465	67.1	362812	224090	143236	79232	89334	645755
Random term	Real donor	176056	66.1	390072	248847	160464	86722	124964	731064

Notations: NN = Nearest Neighbour, 95% = 95% Quantile, and respectively for 75% and 25%, MD = Median, MAE = Mean Absolute Error. MAXE = Maximum Error.
Explanations: Linear Regression is made using SOLAS 3.0; SOM and MLP are made using NDAEI; Random hot decking (random draw) and RBNN are made by SAS.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 1037–1050

ASYMPTOTIC CONSIDERATIONS CONCERNING REAL TIME SAMPLING METHODS

Kadri Meister[1]

ABSTRACT

In this paper some asymptotic results concerning real time sampling methods are discussed. The population is passing a sampler in real time and for every unit the sampler decides immediately whether or not to sample it. Alternatively, the sampler is passing the population. For such a sampling design, the asymptotic model–based expectation of the mean square error (MSE) of the sample mean is studied. It depends on both population model correlations and sampling correlations. We are interested in the possible gain in efficiency when using real time sampling designs with negative sampling correlations compared with Bernoulli sampling. Under certain conditions optimal sampling correlation have a simple form. In the best case the MSE under a design with negative sampling correlations is around one third of that for the Bernoulli design. However, only if the population correlations are close to 1, this gain is attainable. Numerical calculations are presented for some stationary autocorrelated population models.

Key words: Real time sampling designs, Bernoulli sampling, negative sampling correlations, autocorrelated populations, mean square error (MSE).

1. Introduction

We look at a finite population and a sampling situation where units come, one by one, in real time to a sampler. For every unit the sampler should decide immediately whether or not to sample it by using some sequential selection method. Alternatively, the sampler visits the units in some order chosen by the sampler. The size N of the population is most often unknown in advance. This kind of sampling is here referred to as *real time sampling*.

Some suitable real time sampling methods are Bernoulli sampling, sampling according to a stationary process, and renewal sampling, where the latter can be described as a sampling with independent random step-lengths. The Bernoulli design is a sampling design, where the sampling correlations are equal to zero.

[1] Department of Mathematical Statistics, Umeå University, SE-90187 Umeå, Sweden.

However, it is often of interest to avoid getting units close to each other sampled too frequently. For the other real time sampling designs, described in Meister and Bondesson (2001), the sampling depends on what has happened in the past. One can sample with negative autocorrelations and get estimates with better properties. The aim of this work is to make comparisons between real time sampling designs with negative auto-correlations and the Bernoulli design.

Let I_1, I_2, ... be indicator variables telling whether or not a unit in the population should be sampled. We consider the case when $\{I_i\}$ is a real stationary Bernoulli process in discrete time. We are then sampling with equal inclusion probabilities.

To estimate the population mean $\overline{Y}$ we use the Horvitz-Thompson (HT) ratio estimator. It reduces to the sample mean in our case. We are interested in properties of the mean square error (MSE) of $\overline{y}$. Some knowledge of the structure of the population is needed. Therefore we use a model–based approach. The population values are no longer assumed to be fixed unknown constants. Instead they are assumed to be realized outcomes of random variables Y_1, Y_2, …, Y_N with an N-dimensional probability distribution. Thus the population values are assumed to be generated by a model, a superpopulation model. The model–based approach is described by, e.g., Särndal, Svensson and Wretman (1992, Chapter 14.5). We use the symbol Y_i to denote both the random variable associated with the ith element and an outcome of this random variable.

We do not want to estimate the parameters of the population model. We use the model to understand in which cases one sampling design is preferable to another. Using a so-called hybrid approach, we consider the model-based expectation of the $MSE(\overline{y})$ under a random sampling scheme.

In Meister and Bondesson (2001) some simulation studies were performed for a rather specific $AR(2)$-model. We were interested in the behaviour of the MSE and MSE–estimates of the sample mean for different real time sampling designs. Due to the main aim to see differences in MSE for different sampling designs, a rather small population size $N = 48$ was considered. Renewal sampling with step-lengths according to a negative binomial distribution was found to be the "winner". Yet, one of the conclusions was that one should look at larger populations in simulation studies.

There are many different population models one would like to consider and a simulation study may be time consuming. Asymptotic calculations for some specified population models would give better insight concerning situations where more sophisticated real time sampling methods would be preferable to Bernoulli sampling.

In general we are interested in finding out which real time sampling design is optimal in combination with the sample mean as the estimate. We use some models with autocorrelation for the population.

Many optimality problems in sampling have been considered and solved. Bellhouse (1984) reviews six different types of optimality problems.

Autocorrelated populations are considered in Cochran (1977, Chapter 8) for comparing the model–based expected variance of $\bar{y}$ in the case of systematic sampling with stratified and simple random sampling, respectively. Of the three designs, the systematic design is optimal when the correlation function is decreasing and convex. This result has been generalized by others. However, systematic sampling has some disadvantages in the real time sampling case. It gives certainly bias if the sampler determines the order of the units or these can order themselves by taking into account the sampling procedure. Also, for systematic sampling there is no correct MSE–estimate.

In the following we restrict our study to properties of $MSE(\bar{y})$. In Section 2 the derivation of the asymptotic expected $MSE(\bar{y})$ is given. The stationary autocorrelated population case is of special interest. In Section 3 the problem of minimum asymptotic expected MSE($\bar{y}$) is considered. A simple method for calculating the best sampling correlations is given. In Section 4 numerical examples are presented for some population models. Sampling designs with negative sampling correlations are compared with the Bernoulli design. In Section 5, some conclusions are given.

2. Asymptotic expected MSE

In this section the HT ratio estimator as an estimator of $\bar{Y}$ is considered. Under certain conditions it reduces to $\bar{y}$. The MSE estimator for $\bar{y}$ is given. The asymptotic model–b expectation of $MSE(\bar{y})$ is derived.

2.1. Sampling and estimation method

Let $U = \{1, 2, \ldots, N\}$ be a finite population with the study variable Y, where the Y–values are generated by some population model ξ. Let $I_i \in \{0, 1\}$ be the inclusion indicator for unit i, where $I_i = 0$ means that unit i is not sampled and $I_i = 1$ means that it is sampled. We assume that the indicator variables $\{I_i\}$ are generated by a stationary Bernoulli(π)–process. For fixed N the sequence $\{I_i\}$ has a multivariate Bernoulli distribution. This distribution can be seen as a sampling design (Traat, 2000).

We are sampling with equal first order inclusion probabilities, $\Pr(I_i = 1) = \pi$ for all i. The second order inclusion probabilities $\pi_{ij} = \Pr(I_i = 1, I_j = 1)$ are equal for all i, j where $|j - i| = k$, $k \neq 0$. The design characteristics are

$$E(I_i) = \pi,$$
$$Var(I_i) = \pi(1-\pi),$$
$$Corr(I_i, I_j) = \pi_{ij} - \pi^2, \quad i \neq j.$$

For varying inclusion probabilities π_i the population mean $\bar{Y}$ is estimated by the ratio estimate

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{N} Y_i I_i/\pi_i}{\sum_{i=1}^{N} I_i/\pi_i} \tag{1}$$

If $\pi_i \equiv \pi$, as in our case, this estimate reduces to the sample mean $\bar{y}$. Since $\{I_i\}$ is a random sequence, n is also random. Obviously $En = N\pi$. Since $\bar{y}$ is a ratio estimate, it may have a slight bias.

Rewriting $\bar{y} - \bar{Y}$ as $\sum_{i=1}^{N}(I_i/n - 1/N)Y_i$, we find that the MSE of $\bar{y}$ is

$$MSE(\bar{y}) = E[(\bar{y} - \bar{Y})^2] = \sum_{i=1}^{N}\sum_{j=1}^{N} E\left[\left(\frac{I_i}{n} - \frac{1}{N}\right)\left(\frac{I_j}{n} - \frac{1}{N}\right)\right] Y_i Y_j$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} E\left[\left(\frac{I_i}{n} - \frac{1}{N}\right)\left(\frac{I_j}{n} - \frac{1}{N}\right)\right](Y_i - Y_j)^2 ,$$

where the last expression is obtained by some simple algebra. The coefficients $a_{ij} = E[(I_i/n - 1/N)(I_j/n - 1/N)]$ can be approximated by

$$\alpha_{ij} \approx [En]^{-2} E\left[\left(I_i - \frac{n}{N}\right)\left(I_j - \frac{n}{N}\right)\right] = [En]^{-2}\left(\pi_{ij} - \bar{\pi}_{i.} - \bar{\pi}_{j.} + \bar{\pi}_{..}\right),$$

where

$$\bar{\pi}_{i.} = \frac{\pi_{i.}}{N} \quad \text{with} \quad \pi_{i.} = \sum_{j=1}^{N}\pi_{ij} = E(I_i n),$$

$$\bar{\pi}_{..} = \frac{\pi_{..}}{N^2} \quad \text{with} \quad \pi_{..} = \sum_{i=1}^{N}\sum_{j=1}^{N}\pi_{ij}.$$

Hence,

$$MSE(\bar{y}) \approx -\frac{1}{2[En]^2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}(Y_i - Y_j)^2 , \tag{2}$$

cf. Meister and Bondesson (2001).

An approximately unbiased estimate of $MSE(\bar{y})$ is

$$\hat{MSE}(\bar{y}) = -\frac{1}{2[En]^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{a_{ij}}{\pi_{ij}}(Y_i - Y_j)^2 I_i I_j , \tag{3}$$

where the factor $1/[En]^2$ can be replaced by $1/n^2$.

2.2. Asymptotic expected MSE for some population models

We look at a population model ξ, that is described by

$$EY_i = \mu, \quad Var(Y_i) = \sigma^2, \quad Corr(Y_i, Y_j) = \rho_{ij}.$$

The model–based expected value of the MSE (2) is

$$E_\xi\left[MSE(\bar{y})\right] \approx -\frac{1}{2[En]^2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} E_\xi\left[(Y_i - Y_j)^2\right]$$

$$= -\frac{\sigma^2}{[En]^2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}(1-\rho_{ij}). \qquad (4)$$

Since $\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} = 0$, (4) reduces to

$$E_\xi\left[MSE(\bar{y})\right] \approx \frac{\sigma^2}{[En]^2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\rho_{ij}. \qquad (5)$$

If N is large, we can assume that $\pi_{ij} \to \pi^2$ when $|j - i|$ increases. This implies that $\overline{\pi}_{i.} \to \pi^2$ and hence

$$a_{ij} \approx \pi_{ij} - \pi^2 = \pi(1-\pi)R_{ij}, \qquad (6)$$

where $R_{ij} = Corr(I_i, I_j)$. Inserting this approximation into (5), we get

$$E_\xi\left[MSE(\bar{y})\right] \propto \sum_{i=1}^{N}\sum_{j=1}^{N} R_{ij}\rho_{ij}.$$

Let now ξ be a stationary population model, described by

$$EY_i = \mu, \quad Var(Y_i) = \sigma^2, \quad Cov(Y_i, Y_{i+k}) = \rho_k\sigma^2,$$

where $\rho_k \geq 0$ is the lag k correlation. Now we can rewrite formula (5), using the approximations (6), as

$$E_\xi\left[MSE(\bar{y})\right] \approx \frac{\sigma^2\pi(1-\pi)N}{[En]^2}\left[1 + 2\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)R_k\rho_k\right], \qquad (7)$$

where $R_k = Corr(I_i, I_{i+k})$ and $En = N\pi$. We focus on this case in the sequel.
For Bernoulli sampling, all R_k values ($k \neq 0$) are equal to zero, because of independence between the inclusion indicators. Formula (7) reduces to the form

$$E_\xi\left[MSE_{BE}\left(\bar{y}\right)\right] \approx \frac{\sigma^2\pi\left(1-\pi\right)}{\left[En\right]^2} N.$$

We look at the ratio

$$\frac{E_\xi\left[MSE_{NC}\left(\bar{y}\right)\right]}{E_\xi\left[MSE_{BE}\left(\bar{y}\right)\right]} \approx 1 + 2\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)R_k\rho_k, \tag{8}$$

where *NC* indicates a sampling design with negative sampling correlations. The ratio expresses how well a sampling design with negative sampling correlations performs compared with the Bernoulli design with the same π. When the value of (8) is less than 1, there are advantages by sampling with negative correlations.

One should try to minimize the value of (8) for a given population model, i.e., try to find the best correlations R_k for the indicator variables.

For many natural populations, one may expect Y_i and Y_j to be more alike when i and j are close together in the series than when they are distant. We assume the following condition for the population model

$$\rho_k \downarrow 0 \quad \text{(not too slowly) as } k \to \infty. \tag{9}$$

Due to the assumption that neighbouring units may have similar *Y*–values, it may be wise not to sample pairs of units close to each other too often. In general, if ρ_k is high, pairs with units lag k apart should be sampled rarely. If ρ_k is small, pairs should be sampled often. Therefore one should try to use *negative sampling correlations* for pairs of units with high ρ_k values.

Under the condition that $\sum_{k=1}^{\infty} R_k\rho_k$ converges, a partial summation shows that for large N the ratio (8) can be approximated by

$$e = 1 + 2\sum_{k=1}^{\infty} R_k\rho_k. \tag{10}$$

Thus one should try to minimize $g = \sum_{k=1}^{\infty} R_k\rho_k$ for given ρ_k values, assumed to be positive.

Remark. One can also compare the variances of the sample sizes. The sample size n is given by $n = \sum_{i=1}^{N} I_i$, and hence

$$Var\left(n\right) = \pi\left(1-\pi\right)N\left[1 + 2\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)R_k\right].$$

For Bernoulli sampling, $Var(n) = \pi(1-\pi)N$. Looking at the approximated ratio of variances we get

$$e_n = 1 + 2\sum_{k=1}^{\infty} R_k .$$

We see that negative sampling correlations lead to lower variance for *n*.

3. Minimization of the asymptotic expected MSE

In this section we try to solve the minimization problem above. It is obvious from (10) that one should try to make the sampling correlations as negative as possible for *k* small. There are some restrictions on the sampling correla-tions for getting stable estimates. However, it turns out that the best negative sampling correlations have a simple form.

3.1. Optimal sampling correlations

We assume that $\{R_k\}$ is a sequence of correlations, such that

$$0 \geq R_k \geq -d, \quad k = 1,2,\ldots, \tag{11}$$

$$\sum_{k=1}^{\infty} R_k \geq -\frac{1}{3}, \tag{12}$$

where $d = (1 - \delta)\pi/(1 - \pi)$ and δ is some suitable constant, $0 < \delta < 1$.

Condition (11) is needed for the stability of the MSE–estimate (3). The values of a_{ij} should preferably be negative to get a stable MSE–estimate, and therefore the second order inclusion probabilities should satisfy $\pi_{ij} \leq \pi^2$. Since π_{ij} is in the denominator of (3), π_{ij} should not be too small as otherwise the terms forming the MSE–estimate will vary a lot. Thus the restriction $\pi_{ij} \geq \delta\pi^2$ should be satisfied. Since R_k depends on the second order inclusion probabilities $\pi_{i,i+k}$ as

$$R_k = \frac{\pi_{i,i+k} - \pi^2}{\pi(1-\pi)},$$

we get the restrictions for the R_ks as in (11). An increase of the value of δ increases the possible negative lower bound for the R_k values.

Condition (12) is due to the fact that the R_ks are correlations for Bernoulli variables. In Bondesson (2000) the following conjecture is given: if $R_k \leq 0$ for $k \geq 1$, then $\sum_{k=1}^{\infty} R_k \geq -1/3$. Condition (12) is based on this simple conjecture. The exact lower bound for $\sum_{k=1}^{\infty} R_k$ is under study. It may be slightly lower than –1/3 but never lower than –1/2. The bound may also depend on π. *For the sake of simplicity we will assume it to be* –1/3.

It follows that we have to solve a linear optimization problem. One has to find correlations R_k, so that the conditions (11) and (12) are fulfilled and $g = \sum_{k=1}^{\infty} R_k \rho_k$ attains its minimum value.

Assume that condition (9) is fulfilled. We should have the absolute values of the correlations R_k, with k small, as large as possible.

PROPOSITION 1. *Assume that it is possible to obtain equality in* (12)*. Then the optimal correlations are as follows.*

$$R_k = \begin{cases} -d & for \quad k < L \\ -1/3 + (L-1)d & for \quad k = L \\ 0 & for \quad k > L \end{cases}, \tag{13}$$

where L is the smallest integer such that $Ld \geq 1/3$.

Proof. Let first $d \geq 1/3$. Then we get $R_1 = -1/3$, $R_k = 0$, $k > 1$, and $g = -\rho_1/3$. This is the minimum value since

$$g = \sum_{k=1}^{\infty} R_k \rho_k \geq \rho_1 \sum_{k=1}^{\infty} R_k \geq -\frac{1}{3}\rho_1 .$$

Let then $d < 1/3$ and let L be as above. Then

$$g = \sum_{k=1}^{\infty} R_k \rho_k \geq \sum_{k=1}^{L-1} R_k \rho_k + \rho_L \sum_{k=L}^{\infty} R_k$$

$$= \sum_{k=1}^{L-1} R_k \rho_k + \rho_L \left(\sum_{k=1}^{\infty} R_k - \sum_{k=1}^{L-1} R_k \right) = \sum_{k=1}^{L-1} R_k (\rho_k - \rho_L) + \rho_L \sum_{k=1}^{\infty} R_k$$

$$\geq \sum_{k=1}^{L-1} (-d)(\rho_k - \rho_L) + \left(-\frac{1}{3} \right) \rho_L = \sum_{k=1}^{L-1} (-d)\rho_k + \left(-\frac{1}{3} + (L-1)d \right) \rho_L .$$

Thus the lower bound is attained by the R_ks given in the proposition.

Since $(L-1)d < 1/3$, it also easily follows by some calculations that

$$\sum_{k=1}^{L-1} (-d)\rho_k + \left(-\frac{1}{3} + (L-1)d \right) \rho_L \geq -\frac{1}{3} \frac{\sum_{k=1}^{L-1} \rho_k}{L-1} .$$

Hence $g \geq -1/3$ and it follows that $e = 1 + 2g \geq 1/3$. The value of e is close to 1/3 only if all the correlations $\rho_1, \ldots, \rho_{L-1}$ are close to 1.

3.2. Possible sampling designs with optimal correlations

The simple structure of the optimal sequence $\{R_k\}$ is interesting. Of course, it is also of great interest to have a sampling method for which it can be implemented.

Renewal sampling according to Sykes' distribution (cf. Bondesson, 1986, and Meister and Bondesson, 2001) has approximately this structure. The step–lengths have a complicated distribution but

$$\Pr(I_{i+k} = 1 \mid I_i = 1) = \begin{cases} a & \text{if} \quad k < L \\ \pi & \text{if} \quad k \geq L \end{cases},$$

where $a \leq \pi$ and

L is some integer. For the existence of such a sampling method, it is necessary and sufficient that

$$\pi \leq a + \frac{(1-a)^L}{L}\left(1 - \frac{1}{L}\right)^{L-1}.$$

It follows that

$$R_k = \begin{cases} (a-\pi)/(1-\pi), & k < L \\ 0, & k \geq L \end{cases}.$$

Sampling according to a stationary process (cf. Meister and Bondesson, 2001) is another method for getting equal sampling correlations R_k. Let $\{Z_i\}$ be a stationary normal process, $Z_i \sim N(0,1)$. Unit i is included in the sample if $Z_i \leq c$, where c is a constant that depends on the predetermined π. Let $r_k = Corr(Z_i, Z_{i+k})$. For a normal process $\{Z_i\}$, the correlation r_k determines the correlation R_k and if r_k is negative so is R_k. A sequence $\{r_k\}_1^\infty$ of negative numbers is autocorrelation sequence for a stationary normal process if and only if $\sum_{k=1}^{\infty} r_k \geq -1/2$ (Bondesson, 2000). Thus, setting

$$r_k = \begin{cases} -\dfrac{1}{2(L-1)}, & k < L \\ 0, & k \geq L \end{cases},$$

we get sampling correlations R_k that are equal for $k < L$ and zero for $k \geq L$. The lower bound for the sampling correlations R_k is $-1/(3(L-1))$ for $k < L$. The bound is almost attained for $\pi = 1/2$.

Whether there is a sampling design with the optimal correlations, depends on the inclusion probabilities. For different sampling designs the equality in (12) is not valid for all π values and sampling with $R_1 = -1/3$ is not possible.

4. Some numerical comparisons

We are interested in seeing how the measure e in (10) behaves under different population models. As always in this type of studies, one meets the problem that there is an "ocean" of situations of potential interests. To restrict ourselves, we look at situations where the population is generated by an $MA(q)$– or $AR(p)$–model. We present numerical values of e for specific such population models.

We let π take the values 0.1, 0.2, 0.3, 0.4, 0.5 and for every such value we find the minimum value of e for δ = 0.1, 0.3, 0.5, 0.7, 0.9.

4.1. Populations generated by $MA(q)$–models

We assume that the Y–values are generated by a real stationary $MA(q)$–process

$$Y_i = \sum_{j=0}^{q} \alpha_j \eta_{i-j} ,$$

where α_j, j = 0, 1, …, q, are constants and the η_is are i.i.d. random variables. The correlations ρ_k are given by the relations

$$\rho_k = \frac{\sum_{j=0}^{q-k} \alpha_j \alpha_{j+k}}{\sum_{j=0}^{q} \alpha_j^2}, \quad k = 0, 1, 2, \ldots,$$

see, e.g., Chatfield (1996, p. 33). We set $\alpha_0 = 1$. The correlations ρ_k "cut off" at lag q, a special feature of an $MA(q)$–process.

We look at 3 different population models

(a) $MA(1)$ with $\rho_1 = 1/2$;

(b) $MA(2)$ with $\rho_1 = \sqrt{2}/2$, $\rho_2 = 1/4$;

(c) $MA(2)$ with $\rho_1 = 2/3$, $\rho_2 = 1/3$.

For the $MA(2)$–models (ρ_1, ρ_2) are chosen to be extreme points of the set of possible pairs of correlations. For the $MA(1)$–model $\rho_1 = 1/2$ is also an extreme point.

Numerical values of e for different parameter values are given in Table 1.

Table 1. Minimum e for MA(1) and MA(2)–models

π		**0.1**			**0.2**			**0.3**			**0.4**			**0.5**	
δ	(*a*)	(*b*)	(*c*)	(*a*)	(*b*)	(*c*)	(*a*)	(*b*)	(*c*)	(*a*)	(*b*)	(*c*)	(*a*)	(*b*)	(*c*)
.1	.90	.81	.80	.78	.63	.63	.67*	.53*	.56*	.67*	.53*	.56*	.67*	.53*	.56*
.3	.92	.85	.84	.83	.67	.66	.70	.56	.58	.67*	.53*	.56*	.67*	.53*	.56*

.5	.94	.89	.89	.88	.76	.75	.79	.64	.63	.67*	.53*	.56*	.67*	.53*	.56*
.7	.97	.94	.93	.93	.86	.85	.87	.75	.74	.80	.65	.64	.70	.56	.58
.9	.99	.98	.98	.98	.95	.95	.96	.92	.91	.93	.87	.87	.90	.81	.80

In the table * indicates cases where e is minimized by $R_1 = -1/3$. In this case e equals 2/3, $(3-\sqrt{2})/3$, and 5/9 for ρ_1 equal to 1/2, $\sqrt{2}/2$, and 2/3, respectively. For higher values of π there is a tendency that e is minimized by just taking $R_1 = -1/3$.

We can see that sampling with negative correlations gives better results for the *MA*(2)–models than for the *MA*(1)–model; this is an effect of the higher population correlation values. When the value of δ is higher, the e–value is slightly lower for Model (b) than for Model (c); this is an effect of the slightly higher value of $\rho_1 + \rho_2$ for Model (b). The maximum gain (= $1 - e$) is around 33% for the *MA*(1)–model and around 47% for the *MA*(2)–models.

4.2. Populations generated by *AR*(*p*)–models

We assume that the Y–values are generated by an $AR(p)$–process

$$Y_i = \sum_{j=1}^{p} \beta_j Y_{i-j} + \varepsilon_i \ ,$$

where the variables ε_i are i.i.d. random variables. The correlations ρ_k satisfy the Yule–Walker equations (see, e.g., Chatfield, 1996, p. 38)

$$\rho_k = \sum_{j=1}^{p} \beta_j \rho_{k-j} . \tag{14}$$

We want the constants β_j to be chosen so that $\rho_k \geq 0$ and $\rho_k \downarrow$.
We look at 2 different population models

a) $AR(1)$ with $\rho_1 = 0.9$, $\rho_k = 0.9^k$, $k \geq 2$;
b) $AR(2)$ with $\rho_1 = 0.9$, $\rho_2 = 0.85$, and ρ_k for $k \geq 3$, are defined by Yule–Walker's formulae (14). (It can be verified that ρ_k is decreasing.)

Numerical values of e for some parameter values are given in Table 2. The optimal R_ks are the same for the two models. The number of possible nonzero correlations, L, depends on the conditions (11)–(12) and is given for all π– values in the third column in Table 2.

Table 2. Minimum e for AR(1) and AR(2)–models

π		**0.1**			**0.2**			**0.3**			**0.4**			**0.5**	
δ	(*a*)	(*b*)	*L*	(*a*)	(*b*)	*L*	(*a*)	(*b*)	*L*	(*a*)	(*b*)	*L*	(*a*)	(*b*)	*L*
.1	.47	.44	4	.42	.41	2	.40	.40	1	.40	.40	1	.40	.40	1

.3	.49	.46	5	.43	.42	2	.41	.40	2	.40	.40	1	.40	.40	1
.5	.53	.49	6	.45	.43	3	.42	.41	2	.40	.40	1	.40	.40	1
.7	.61	.54	10	.50	.46	5	.45	.43	3	.42	.41	2	.41	.40	2
.9	.81	.73	30	.66	.59	14	.57	.51	8	.51	.47	5	.47	.44	4

The optimal correlations are $R_k = -d$, $k < L$, $R_L = -1/3 + (L - 1)d$, and $R_k = 0$, $k > L$, where L is the smallest integer such that $Ld \geq 1/3$. For example, for $\pi = 0.1$ and $\delta = 0.5$, the minimum e is attained when $L = 6$.

The e–value is never below 0.4. We see from the L–column that for fixed π, we should let more correlations be negative when δ increases to get a low e–value. When we compare the e–values for the $AR(1)$– and $AR(2)$–models, we see that they are somewhat lower for the $AR(2)$–model. This is due to fact that the correlations for this model are higher than for the $AR(1)$–model.

Table 3. Values of minimum e for AR(1)–model

π		**0.1**			**0.2**			**0.3**			**0.4**			**0.5**	
		ρ_1			ρ_1			ρ_1			ρ_1			ρ_1	
δ	0.9	0.4	0.1	0.9	0.4	0.1	0.9	0.4	0.1	0.9	0.4	0.1	0.9	0.4	0.1
.1	.47	.87	.98	.42	.79	.95	.40	.73	.93	.40	.73	.93	.40	.73	.93
.3	.49	.90	.98	.43	.81	.96	.41	.75	.94	.40	.73	.93	.40	.73	.93
.5	.53	.93	.99	.45	.85	.97	.42	.79	.95	.40	.73	.93	.40	.73	.93
.7	.61	.96	.99	.50	.90	.98	.45	.85	.97	.42	.80	.96	.41	.75	.94
.9	.81	.99	1.0	.66	.97	.99	.57	.94	.99	.51	.91	.99	.47	.87	.98

Sampling with negative correlations gives advantages compared with Bernoulli sampling when the population correlation ρ_1 is relatively large. In Table 3 the values of e are given for the $AR(1)$–model when ρ_1 equals 0.9, 0.4, and 0.1. For $\rho_1 = 0.9$ the best result is obtained. The maximum possible gain is 60% compared with Bernoulli sampling. For $\rho_1 = 0.4$ the gain is approximately 30%. However, the sampling designs give similar results for small values of π. For $\rho_1 = 0.1$, a gain less than 10% is obtained, so in this case it is advantageous to use Bernoulli sampling due to its simplicity.

Different sampling methods can attain maximum possible gain, depending on the inclusion probabilities π. For $\pi = 0.5$, sampling according to a stationary normal process with $r_1 = -1/2$, gives $R_1 = -1/3$, but for $\pi = 0.4$ the lower bound for R_1 is –0.262. For renewal sampling according to Sykes' distribution the lower bound for R_1 is –0.171 and –0.225, respectively.

5. Conclusions

For a stationary population model with decreasing correlations, we conclude that sampling with negative correlations gives smaller asymptotic model–based expectation of $MSE(\bar{y})$ than Bernoulli sampling. A sampling design with correlations $R_k = -d$ for $k < L$ and $R_k = 0$ otherwise, where L is some integer and d an appropriate constant, is approximately optimal. There are several sampling designs with correlations of the desired form. However, it is not always possible to attain the correlations suggested in Proposition 1. Further research about suitable sampling methods is needed.

Acknowledgements

I am grateful to Lennart Bondesson and Imbi Traat for valuable and encouraging comments on this paper. Additional comments from a referee have improved the presentation. This work was partly supported by a scholarship of the New Visby Programme of the Swedish Institute.

REFERENCES

BELLHOUSE, D. R. (1984), A review of optimal designs designs in survey sampling. *Canadian Journal of Statistics* 12, 53-65.

BONDESSON, L. (1986), Sampling of a linearly ordered population by selection of units at successive random distances. *Report No. 25*, Section of Forest Biometry, Swedish University of Agricultural Sciences, Umeå, Sweden.

BONDESSON, L. (2000), On a Minimum Correlation Problem. *Research Report No. 9*, Department of Mathematical Statistics, Umeå University, Umeå, Sweden.

CHATFIELD, C. (1996), *The Analysis of Time Series. An Introduction.* 5th ed. London: Chapman & Hall.

COCHRAN, W. G. (1977), *Sampling Techniques*. 3rd ed. New York: Wiley.

MEISTER, K. and BONDESSON, L. (2001), Some Real Time Sampling Methods. *Research Report No. 2*, Department of Mathematical Statistics, Umeå University, Umeå, Sweden.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

TRAAT, I. (2000), Sampling design as a multivariate distribution. In

T. Kollo, E.-M. Tiit, M. Srivastava (Ed-s). *New trends in Probability and Statistics 5, Multivariate Statistics*. Vilnius, Utrecht: TEV/VSP, 195-208.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 1051–1066

MODIFIED CHAIN RATIO ESTIMATORS FOR FINITE POPULATION MEAN USING TWO AUXILIARY VARIABLES IN DOUBLE SAMPLING

B. Prasad[1], Radhey S. Singh[2] and Housila P. Singh[3]

ABSTRACT

This paper proposes modified ratio estimators for finite population mean using two auxiliary variables in double sampling. Asymptotic expressions for biases and mean square errors of the proposed estimators are obtained. Asymptotic optimum estimators (AOEs) are also identified. Further, the optimum values (depending upon population parameters) when replaced from sample values yields the estimators having the mean square errors of the AOEs. An empirical study is carried out to demonstrate the performances of the constructed estimators over conventional unbiased estimator, traditional ratio estimator in double sampling and the estimator suggested by Chand(1975), Kiregyera(1980) and Upadhyaya et. al.(1990).

Key words and Phrases **:** Finite population mean, Chain ratio estimator, Double sampling, Auxiliary variables, Asymptotic optimum estimator, Bias, Mean square error.

1. Introduction

Let $U = \left(U_1, U_2, \ldots, U_N\right)$ be a finite population of N (given) units. Let y and (x,z) denote the study and auxiliary variates taking the values y_i and $\left(x_i, z_i\right)$ on the unit $U_i\,(i = 1,2,\ldots,N)$.

It is well known that in most of the survey situations, auxiliary information is available (or may be made to be available diverting some of the resources) in one or the other form. If used intelligibly, this information may provide us the estimators better that those in which no auxiliary information is used.

[1] Statistics Canada, Canada.
[2] University of Guelph, Canada.
[3] School of Studies in Statistics, Vikram University, Ujjain-456 010, INDIA.

Let $y_1, y_2, \ldots, y_n$ be a simple random sample of size *n* drawn without replacement from *U*. Then to estimate the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ of the study variable *y*, the traditional ratio estimator based on a simple random sample of size *n*, is defined by

$$\hat{\overline{Y}}_R = \overline{y}\left(\overline{X} / \overline{x}\right) \tag{1.1}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$, the population mean of auxiliary variable *x* is known.

It is well known result (e.g. Cochran(1977)) that $\hat{\overline{Y}}_R$ will estimate $\overline{Y}$ to terms of order O(1/*n*) more precisely $\overline{y}$ if $\rho_{yx} > (1/2)\left(C_x / C_y\right)$, where ρ_{yx} is the correlation coefficient between *y* and *x*, and C_x, C_y are coefficients of variations of *x*, *y* respectively.

In many situations of practical importance, the population mean $\overline{X}$ of the auxiliary variates *x* is not known before start of the survey. The usual practice is to estimate it by the sample mean $\overline{x}' = \left(1/n'\right)\sum_{i=1}^{n'} x_i$, where n' is the size of a preliminary simple random sample drawn from *U* without replacement and $n < n'$ is a sub sample of n'. Then the ratio estimator in double sampling is defined by

$$\hat{\overline{Y}}_{Rd} = \overline{y}\left(\overline{x}'/\overline{x}\right) \tag{1.2}$$

Suppose that $\overline{Z} = \left(1/N\right)\sum_{i=1}^{N} z_i$, the population mean of another auxiliary variate z closely related to x but compared to x remotely related to y is available (e.g. y is the value of cattle and/or calves sold live in 1964, x is the number of cattle and/or calves sold in 1964 and z is the number of farms reporting sale of cattle and/or calves sold live in 1964, $\rho_{yx} > \rho_{yz}$; ρ_{yz} is the correlation coefficient between y and z). This type of situation has been briefly discussed by, among others, Kiregyera(1980), Sahoo and Swain(1983), Srivastava et. al(1989,90) and Srivenkataramana and Tracy(1989). Thus, the ratio estimator

$$\hat{\overline{X}}_R = \overline{x}'\left(\overline{Z}/\overline{z}'\right) \tag{1.3}$$

will estimate $\bar{X}$ to terms of order O(1/*n*) more efficiently than $\bar{x}'$ if $\rho_{xz} > (1/2)(C_z/C_x)$,where ρ_{xz} is the correlation coefficient between x and z, and C_z is the coefficient of variation of z.

Replacing $\bar{x}'$ in (1.2) by $\hat{\bar{X}}_{Rd}$, Chand(1975) suggested a chain ratio-type estimator for $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(c)} = (\bar{y}/\bar{x})\hat{\bar{X}}_{Rd} \tag{1.4}$$

$$\hat{\bar{Y}}_{Rd}^{(c)} = \bar{y}(\bar{x}'/\bar{x})(\bar{Z}/\bar{z}')$$

Kiregyera(1980) suggested a chain ratio-to-regression estimator for $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(k)} = (\bar{y}/\bar{x})\left[\bar{x}' + b_{xz}(\bar{Z} - \bar{z}')\right] \tag{1.5}$$

b_{xz} being the estimate of regression coefficient x on z.

Modified this Upadhyaya et. al(1990) suggested a chain ratio-type estimator for $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(u)} = \bar{y}\frac{\left[\bar{x}' + b_{xz}(\bar{Z} - \bar{z}')\right]}{\left[\bar{x}' + \hat{C}(\bar{x} - \bar{x}')\right]} \tag{1.6}$$

where $\hat{C} = (\bar{x}/\bar{y})b_{yx}$, b_{yx} being the estimate of regression coefficient of y on x.

In this paper three chain ratio-type estimators, one based on Srivevkataramana and Tracy(1980,81) transformation and other two based on Walsh(1970) and Reddy(1974) approach, have been proposed and their properties are studied for large samples.

2.1. The Proposed Chain Ratio-Type Estimator

Using the transformation $U_i = A - Z_i, i = 1,2,\ldots\ldots,N$,(A, being suitably chosen scalar) suggested by Srivenkataramana and Tracy(1980,81), we suggest an alternative chain ratio-type estimator for $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(A)} = \bar{y}(\bar{x}'/\bar{x})(\bar{u}/\bar{U}) \tag{2.1}$$

where $\bar{u}' = A - \bar{Z}'$ such that $E(\bar{u}') = \bar{U} = A - \bar{Z}$

To obtain the bias and mean square error(MSE) of $\hat{\bar{Y}}_{Rd}^{(A)}$, we write

$$\bar{y} = \bar{Y}\left(1+e_0\right), \bar{x} = \bar{X}\left(1+e_1\right), \bar{x}' = \bar{X}\left(1+e_1'\right), \bar{z}' = \bar{Z}\left(1+e_2'\right)$$

such that

$$E\left(e_0\right) = E\left(e_1\right) = E\left(e_1'\right) = E\left(e_2'\right) = 0$$

$$E\left(e_0^2\right) = \left(f/n\right)C_y^2, E\left(e_1^2\right) = \left(f/n\right)C_x^2$$

$$E\left(e_2'\right) = \left(f'/n'\right)C_z^2, E\left(e_0, e_1\right) = \left(f/n\right)\rho_{yx}C_yC_x, E\left(e_0, e_1'\right) = \left(f'/n'\right)\rho_{yx}C_yC_x,$$

$$E\left(e_0, e_2'\right) = \left(f'/n'\right)\rho_{yz}C_yC_z$$

$$E\left(e_0, e_2'\right) = E\left(e_1', e_2'\right) = \left(f'/n'\right)\rho_{xz}C_xC_z \tag{2.2}$$

where f= $(N-n)/N$ and $f' = (N-n')/N$.

Expressing (2.1) in terms of e's we have

$$\hat{\bar{Y}}_{Rd}^{(A)} = \bar{Y}\left(1+e_0\right)\left(1+e_1'\right)\left(1+e_1\right)^{-1}\left(1-\theta e_2'\right) \tag{2.3}$$

where $\theta = \bar{Z}/\left(A-\bar{Z}\right)$.

Assuming $\left|e_1\right| < 1$,expanding the right hand side of (2.3) and retaining terms up to second degree of e's, we have

$$\left(\hat{\bar{Y}}_{Rd}^{(A)} - \bar{Y}\right) = \bar{Y}\left[e_0 - e_1 + e_1' - \theta e_2' - e_0e_1 + e_0e_1' - e_1e_1' - \theta\left(e_0e_2' + e_1'e_2' - e_1e_2'\right) + e_1^2\right] \tag{2.4}$$

Taking expectation both sides of (2.4), we get the bias of $\hat{\bar{Y}}_{Rd}^{(A)}$ to the first degree of approximation (or alternatively), to terms of order O(1/n) as

$$B\left(\hat{\bar{Y}}_{Rd}^{(A)}\right) = \bar{Y}\left[\lambda C_x^2(1-C) - \left(f'/n'\right)\theta C^*C_z^2\right] \tag{2.5}$$

where $\lambda = \left(\frac{1}{n} - \frac{1}{n'}\right), C = \rho_{yx}\left(C_y/C_x\right)$ and $C^* = \rho_{yz}\left(C_y/C_z\right)$.

Squaring both sides of (2.4), retaining terms of e's up to second degree and taking expectation, we get the mean square error of $\hat{\bar{Y}}_{Rd}^{(A)}$ to the first degree of approximation, as

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(A)}\right) = \bar{Y}^2\left[\left(f/n\right)C_y^2 + \lambda C_x^2(1-2C) + \left(f'/n'\right)C_z^2\theta\left(\theta - 2C^*\right)\right] \quad (2.6)$$

which is minimized for

$$\theta = C^* = \theta_0 \text{ (say)}$$

$$\Rightarrow \qquad A = \left(\frac{1+C^*}{C^*}\right)\bar{Z} = A_0 \text{ (say)} \quad (2.7)$$

Putting (2.7) in (2.6), we get the minimum MSE of $\hat{\bar{Y}}_{Rd}^{(A)}$ as

$$\min MSE\left(\hat{\bar{Y}}_{Rd}^{(A)}\right) = \bar{Y}^2\left[\left(f/n\right)C_y^2 + \lambda C_x^2(1-2C) - \left(f'/n'\right)C_z^2C^{*2}\right] \quad (2.8)$$

Substitution of (2.7) in (2.1) yields the 'asymptotic optimum estimator' (AOE) of $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(A_0)} = \bar{y}\left(\bar{x}'/\bar{x}\right)\frac{\left[\bar{Z} + C^*\left(\bar{Z} - \bar{z}'\right)\right]}{\bar{Z}} \quad (2.9)$$

with same mean square error as given in (2.8).

2.2. Estimator Based on Estimated Optimum

It is to be mentioned that the estimator $\hat{\bar{Y}}_{Rd}^{(A_0)}$ in (2.9) requires the prior knowledge of C^* which is the function of ρ_{yz}, C_y and C_z. In practical sample surveys, a prior value of ρ_{yz}, C_y or C_z and C^* can be guessed quite accurately by utilising appropriate information from a most recent survey taken in the past or by conducting a preliminary survey utilising a small fraction of the full budget allocated for the current survey. For further discussion on this subject the reader is referred to Murthy(1967,pp. 96-99), Reddy(1974) and Sahai and Sahai(1985).

Further, if the investigator is unable to guess the value of C^* the only alternative left to him is to replace C^* in (2.9) by its consistent estimate $\hat{C}^*$ computed from the data at hand. Thus the estimator based on estimated optimum is

$$\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\frac{\left[\bar{Z} + \hat{C}^*\left(\bar{Z} - \bar{z}'\right)\right]}{\bar{Z}} \tag{2.10}$$

where $\hat{C}^* = \left(\bar{Z}/\bar{y}\right)b_{yz}, b_{yz}$ is the estimate of regression coefficient of y on z. To obtain the MSE of $\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}$, we write $\hat{C}^* = C^*\left(1+e_3\right)$ with $E\left(\hat{C}^*\right) = C^* + O(n^{-1})$ and assume that $\left|e_1\right| < 1$.

Now, expressing $\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}$ in terms of e's we have

$$\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)} = \bar{Y}\left(1+e_0\right)\left(1+e_1'\right)\left(1+e_1\right)^{-1}\left\{1 - C^*\left(1+e_3\right)e_2'\right\}$$

where e_0, e_1, e_1', e_2' are same as defined in section (2.1). The MSE of $\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}$ is

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right) = \bar{Y}^2 E\left[\left(1+e_0\right)\left(1+e_1'\right)\left(1+e_1\right)^{-1}\left\{1 - C^*\left(1+e_3\right)e_2'\right\} - 1\right]^2 \tag{2.11}$$

Expanding the right hand side (2.11) and neglecting those terms involving powers of e's greater than two, we have

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right) = \bar{Y}^2 E\left[e_0^2 + e_1^2 + e_1'^2 + C^{*2}e_2'^2 - 2e_0e_1 + 2e_0e_1' - 2C^*e_0e_2' - 2e_1e_1' + 2C^*e_1e_2' - 2C^*e_1'e_2'\right]$$

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right) = \bar{Y}^2\left[\left(f/n\right)C_y^2 + \lambda C_x^2(1-2C) - \left(f'/n'\right)C_z^2C^{*2}\right] \tag{2.12}$$

which is same as given in (2.8).

Thus we proved the following theorem.

Theorem 2.1: The estimator $\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}$ based on estimated optimum has the same MSE to the first degree of approximation as that of AOE $\hat{\bar{Y}}_{Rd}^{(A_0)}$.

2.3. Efficiency Comparison of $\hat{\bar{Y}}_{Rd}^{(A)}$

To compare the estimator $\hat{\bar{Y}}_{Rd}^{(A)}$ with the estimators $\bar{y}, \hat{\bar{Y}}_{Rd}, \hat{\bar{Y}}_{Rd}^{(c)}, \hat{\bar{Y}}_{Rd}^{(k)}$ *and* $\hat{\bar{Y}}_{Rd}^{(u)}$, we write the variance/MSE's of these estimators as

$$V(\bar{y}) = (f/n)\,\bar{Y}^2 C_y^2 \tag{2.13}$$

and to the first degree of approximation,

$$MSE\left(\hat{\bar{Y}}_{Rd}\right) = \bar{Y}^2\left[(f/n)C_y^2 + \lambda C_x^2(1-2C)\right] \tag{2.14}$$

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(c)}\right) = MSE\left(\hat{\bar{Y}}_{Rd}\right) + (f'/n')\bar{Y}^2 C_z^2(1-2C^*) \tag{2.15}$$

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(k)}\right) = MSE\left(\hat{\bar{Y}}_{Rd}\right) + (f'/n')\bar{Y}^2 C_z^2\, C^{**}(C^{**}-2C^*) \tag{2.16}$$

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(u)}\right) = MSE\left(\hat{\bar{Y}}_{Rd}\right) + \bar{Y}^2\left[(f'/n')C_z^2 C^{**}(C^{**}-2C^*) - \lambda C_x^2(1-C)^2\right] \tag{2.17}$$

where $C^{**} = \rho_{xz}\left(C_x/C_z\right)$

For the proof of these results the reader is referred to Kiregyera(1980) and Upadhyaya et al (1990).

When good (specific) guess of C^* is not available we may still have some information about the range of C^*. Such information may be used to provide estimators better than $\bar{y}, \hat{\bar{Y}}_{Rd,}\ \hat{\bar{Y}}_{Rd}^{(c)}, \hat{\bar{Y}}_{Rd}^{(k)}$ and $\hat{\bar{Y}}_{Rd}^{(u)}$.

The results are summarized in the form of theorems. Proof of the theorems are simple so omitted.

Theorem 2.2: The estimator $\hat{\bar{Y}}_{Rd}^{(A)}$ will be better than $\bar{y}$ if $C > 1/2$ and

$$\left.\begin{array}{l}\text{either} \quad 0 < \theta < 2C^* \\ \text{or} \quad\ 2C^* < \theta < 0\end{array}\right\} \tag{2.18}$$

Theorem 2.3: The estimator $\hat{\bar{Y}}_{Rd}^{(A)}$ will dominate over $\hat{\bar{Y}}_{Rd}$ if

$$\left.\begin{array}{l}\text{either} \quad 0 < \theta < 2C^* \\ \text{or} \quad\ 2C^* < \theta < 0\end{array}\right\} \tag{2.19}$$

Theorem 2.4: The estimator $\hat{\bar{Y}}_{Rd}^{(A)}$ will be more efficient than Chand's (1975) estimator $\hat{\bar{Y}}_{Rd}^{(c)}$ if

$$\left.\begin{array}{ll} \text{either} & \left(2C^{*}-1\right)<\theta<1 \\ \text{or} & 1<\theta<\left(2C^{*}-1\right) \end{array}\right\} \tag{2.20}$$

Theorem 2.5: The estimator $\hat{\bar{Y}}_{Rd}^{(A)}$ will be more precised than Kiregyera's (1980) estimator $\hat{\bar{Y}}_{Rd}^{(k)}$ if

$$\left.\begin{array}{ll} \text{either} & \left(2C^{*}-C^{**}\right)<\theta<C^{**};C^{*}<C^{**} \\ \text{or} & C^{**}<\theta<\left(2C^{*}-C^{**}\right);\quad C^{*}>C^{**} \end{array}\right\} \tag{2.21}$$

Theorem 2.6 : The estimator $\hat{\bar{Y}}_{Rd}^{(u)}$ will be better than Upadhyaya et al (1990) estimator if

$$\left(\theta^{2}-2\theta C^{*}\right)<\left[C^{**}\left(C^{**}-2C^{*}\right)-\left(\frac{\lambda n'}{f'}\right)\frac{C_x^{2}(1-C)^{2}}{C_z^{2}}\right] \tag{2.22}$$

2.4. Efficiency Comparison of $\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}$

We have from (2.12) and (2.13) to (2.16) that

$$V(\bar{y})-MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right)=\bar{Y}^{2}\left[\lambda C_x^{2}(2C-1)+\left(f'/n'\right)C_z^{2}C^{*2}\right] \tag{2.23}$$

The condition C>1/2 usually met in survey situations.

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(c)}\right)-MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right)=\bar{Y}^{2}\left(f'/n'\right)C_z^{2}\left(1-C^{*}\right)^{2}>0 \text{ provided } C^{*}\neq 1, \tag{2.24}$$

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(k)}\right)-MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right)=\bar{Y}^{2}\left(f'/n'\right)C_z^{2}\left(C^{**}-C^{*}\right)^{2}>0 \text{ provided } C^{*}\neq C^{**} \tag{2.25}$$

It follows from above expressions that the proposed estimator $\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}$ is better than conventional unbiased estimator $\bar{y}$ with C>1/2, usual ratio estimator $\hat{\bar{Y}}_{Rd}$ in double sampling, Chand's(1975) estimator $\hat{\bar{Y}}_{Rd}^{(c)}$ and the estimator $\hat{\bar{Y}}_{Rd}^{(k)}$ suggested by Kiregyera(1980).

Further we have from (2.12) and (2.17) that

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{\mathrm{Rd}}^{(u)}\right)-\mathrm{MSE}\left(\hat{\bar{Y}}_{\mathrm{Rd}}^{(\hat{A}_0)}\right)=\bar{Y}^2\left[\left(f'/n'\right)C_z^2\left(C^{**}-C^*\right)^2-\lambda C_x^2(1-C)^2\right]$$

which is positive if

$$\left(C^{**}-C^*\right)^2>\left(\frac{\lambda n'}{f'}\right)\left[\frac{C_x}{C_z}(1-C)\right]^2 \tag{2.26}$$

Thus the estimator $\hat{\bar{Y}}_{\mathrm{Rd}}^{(\hat{A}_0)}$ is better than Upadhyaya et al (1990) estimator $\hat{\bar{Y}}_{\mathrm{Rd}}^{(u)}$ provided the condition (2.26) holds good.

3. An Improved Chain Ratio-Type Estimator

Motivated by Walsh (1970) and Reddy (1974), we suggest a class of chain ratio-type estimator for $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(\alpha,\beta)}=\bar{y}\frac{\bar{x}'}{\left[\alpha\bar{x}+(1-\alpha)\bar{x}'\right]}\frac{\bar{Z}}{\left[\beta\bar{z}'+\left(1-\beta\right)\bar{Z}\right]} \tag{3.1}$$

where (α,β) are suitably chosen constants. For (α,β)=(0,0),(1,0),(1,1), $\hat{\bar{Y}}_{\mathrm{Rd}}^{(\alpha,\beta)}$ respectively reduce to $\bar{y}$, $\hat{\bar{Y}}_{\mathrm{Rd}}$ and $\hat{\bar{Y}}_{\mathrm{Rd}}^{(c)}$.

Proceeding as earlier in Section 2.1, the bias and MSE of $\hat{\bar{Y}}_{\mathrm{Rd}}^{(\alpha,\beta)}$, to the first degree of approximation, are respectively given by

$$B\left(\hat{\bar{Y}}^{(\alpha,\beta)}\right)=\bar{Y}\left[\lambda C_x^2\alpha(\alpha-C)+\left(f'/n'\right)C_z^2\beta\left(\beta-C^*\right)\right] \tag{3.2}$$

$$MSE\left(\hat{\bar{Y}}^{(\alpha,\beta)}\right)=\bar{Y}^2\left[\left(f'/n'\right)C_y^2+\lambda C_x^2\alpha(\alpha-2C)+\left(f'/n'\right)C_z^2\beta\left(\beta-2C^*\right)\right] \tag{3.3}$$

The MSE of $\hat{\bar{Y}}_{\mathrm{Rd}}^{(\alpha,\beta)}$ is minimized for

$$\alpha=C=\alpha_0 \text{ (say)}$$

$$\beta=C^*=\beta_0 \text{ (say)} \tag{3.4}$$

Substitution of (3.4) in (3.2) and (3.3) yield respectively the resulting bias and MSE of $\hat{\bar{Y}}_{Rd}^{(\alpha,\beta)}$ as

$$B\left(\hat{\bar{Y}}_{Rd}^{(\alpha,\beta)}\right) = 0$$

$$\min. MSE\left(\hat{\bar{Y}}_{Rd}^{(\alpha,\beta)}\right) = \bar{Y}^2\left[\left(f/n\right)C_y^2 - \lambda C_x^2 C^2 - \left(f'/n'\right)C_z^2 C^{*2}\right] \tag{3.5}$$

Putting (3.4) in (3.1), we get the 'asymptotic optimum estimator' (AOE) as

$$\hat{\bar{Y}}_{Rd}^{(\alpha_0,\beta_0)} = \bar{y}\left(\frac{\bar{x}'}{\bar{x} + C(\bar{x} - \bar{x}')}\right)\left(\frac{\bar{Z}}{\bar{z}' + C^*(\bar{z}' - \bar{Z})}\right) \tag{3.6}$$

which is unbiased to the first degree of approximation and with the same MSE as given in (3.6).

If C and C^* are not known, then replacing C and C^* by their consistent estimates $\hat{C}$ and $\hat{C}^*$ (estimated from the data at hand) in (3.6) we get the estimator based on estimated optimum as

$$\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)} = \bar{y}\frac{\bar{x}'}{\left[\bar{x} + \hat{C}(\bar{x} - \bar{x}')\right]}\frac{\bar{Z}}{\left[\bar{z}' + \hat{C}^*(\bar{z}' - \bar{Z})\right]} \tag{3.7}$$

where $\hat{C} = \left(\bar{x}/\bar{y}\right)b_{yx}, \hat{C}^* = \left(\bar{Z}/\bar{y}\right)b_{yz}, b_{yx}$ and b_{yz} being the estimates regression coefficients of y on x and y on z respectively.

Proceeding as in Section 2.2, it can easily be shown to the first order of approximation that

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}\right) = \bar{Y}^2\left[\left(f/n\right)C_y^2 - \lambda C_x^2 C^2 - \left(f'/n'\right)C_z^2 C^{*2}\right] \tag{3.8}$$

$$= \min MSE\left(\hat{\bar{Y}}_{Rd}^{(\alpha,\beta)}\right) = MSE\left(\hat{\bar{Y}}_{Rd}^{(\alpha_0,\beta_0)}\right)$$

Now, we have from (2.8),(2.17) and (3.8) that

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(u)}\right) - MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}\right) = \left(f'/n'\right)\bar{Y}^2 C_Z^2\left(C^* - C^{**}\right)^2 > 0 \text{ provided } C^* \neq C^{**} \tag{3.9}$$

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right) - MSE\left(\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}\right) = \lambda\bar{Y}^2(1-C)^2 C_x^2 > 0 \text{ provided } C \neq 1 \tag{3.10}$$

It follows from above expressions that $\left(\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}\right)$ is more efficient than Upadhyaya et. al (1990) estimator $\left(\hat{\bar{Y}}_{Rd}^{(u)}\right)$ and the estimator $\left(\hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}\right)$.

A More General Improved Chain Ratio-Type Estimator

Assuming that the data is available for all the units of z-variate (i.e. zi's;i=1,2, …..,n) are known and (n <ń < N) and using the approach of Walsh(1970) and Reddy(1974), we define the following class of chain ratio-type estimator for $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(\alpha,\beta,\gamma)} = \bar{y}\frac{\bar{x}'}{\left[\alpha\bar{x}+(1-\alpha)\bar{x}'\right]}\frac{\bar{Z}}{\left[\beta\bar{z}'+(1-\beta)\bar{Z}\right]}\frac{\bar{Z}}{\left[\gamma\bar{z}+(1-\gamma)\bar{Z}\right]} \tag{4.1}$$

(α,β,γ) being suitably chosen constants and $\bar{z} = (1/n)\sum_{i=1}^{n} z_i$.

The bias and MSE of $\hat{\bar{Y}}_{Rd}^{(\alpha,\beta,\gamma)}$ to the first degree of approximation, are respectively given by

$$B\left(\hat{\bar{Y}}_{Rd}^{(\alpha,\beta,\gamma)}\right) = \bar{Y}\left[\lambda C_x^2\alpha(\alpha-C)+(f'/n')C_z^2\beta(\beta-C^*)-\gamma(f/n)C_z^2C^*+\gamma C_z^2\left\{(f'/n')\beta+\alpha\lambda C^{**}\right\}\right] \tag{4.2}$$

$$MSE\left(\hat{\bar{Y}}_{Rd}^{(\alpha,\beta,\gamma)}\right) = \bar{Y}^2\left[(f/n)\left\{\gamma C_z^2(\gamma-2C^*)+C_y^2\right\}+(f'/n')(\beta+2\gamma-2C^*)\beta C_z^2 + \lambda\alpha\left\{(\alpha-2C)C_x^2+2\gamma C^{**}C_z^2\right\}\right] \tag{4.3}$$

The MSE$\left(\hat{\bar{Y}}_{Rd}^{(\alpha,\beta,\gamma)}\right)$ is minimized for

$$\alpha = \left(\frac{C_y}{C_x}\right)\frac{\left(\rho_{yx}-\rho_{yz}\rho_{xz}\right)}{\left(1-\rho_{xz}^2\right)} = \alpha_0^* \quad \text{(say)}$$

$$\beta = \left(\frac{C_y}{C_z}\right)\frac{\rho_{xz}\left(\rho_{yx}-\rho_{yz}\rho_{xz}\right)}{\left(1-\rho_{xz}^2\right)} = \beta_0^* \quad \text{(say)} \tag{4.4}$$

$$\gamma = \left(\frac{C_y}{C_z}\right)\frac{\left(\rho_{yz}-\rho_{yx}\rho_{xz}\right)}{\left(1-\rho_{xz}^2\right)} = \gamma_0^* \quad \text{(say)}$$

Putting (4.4) in (4.3) we get the minimum MSE of $\hat{\bar{Y}}_{\text{Rd}}^{(\alpha,\beta,\gamma)}$ as

$$\text{MSE}\left(\hat{\bar{Y}}_{Rd}^{(\alpha,\beta,\gamma)}\right) = S_y^2\left[\left(f/n\right) - \lambda R_{y.xz}^2 - \left(f'/n'\right)\rho_{yz}^2\right] \tag{4.5}$$

where

$$S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2, R_{y.xz}^2 = \left(\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}\right)\Big/\left(1-\rho_{xz}^2\right)$$

is the multiple correlation coefficient y on (x,z).

Substitution of (4.4) in (4.1) gives the 'asymptotic optimum estimator' as

$$\hat{\bar{Y}}_{Rd}^{(\alpha_0^*,\beta_0^*,\gamma_0^*)} = \bar{y}\frac{\bar{x}'}{\left[\bar{x}' + \alpha_0^*(\bar{x}-\bar{x}')\right]}\frac{\bar{Z}}{\left[\bar{Z} + \beta_0^*(\bar{z}'-\bar{Z})\right]}\frac{\bar{Z}}{\left[\bar{Z} + \gamma_0^*(\bar{z}-\bar{Z})\right]} \tag{4.6}$$

with MSE same as given in (4.5).

If the optimum values $\left(\alpha_0^*, \beta_0^*, \gamma_0^*\right)$ are not known, replacing $\left(\alpha_0^*, \beta_0^*, \gamma_0^*\right)$ by their consistent estimates

$$\hat{\alpha}_0^* = \left(\frac{\hat{C}_y}{\hat{C}_x}\right)\frac{\left(\gamma_{yx} - \gamma_{yz}\gamma_{xz}\right)}{\left(1-\gamma_{xz}^2\right)}$$

$$\hat{\beta}_0^* = \left(\frac{\hat{C}_y}{\hat{C}_z}\right)\frac{\gamma_{xz}\left(\gamma_{yx} - \gamma_{yz}\gamma_{xz}\right)}{\left(1-\gamma_{xz}^2\right)} \tag{4.7}$$

$$\hat{\gamma}_0^* = \left(\frac{\hat{C}_y}{\hat{C}_z}\right)\frac{\left(\gamma_{yz} - \gamma_{yx}\gamma_{xz}\right)}{\left(1-\gamma_{xz}^2\right)}$$

in (4.6) we get the estimator (based on estimated optimum) for $\bar{Y}$ as

$$\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0^*,\hat{\beta}_0^*,\hat{\gamma}_0^*)} = \bar{y}\frac{\bar{x}'}{\left[\bar{x}' + \hat{\alpha}_0^*(\bar{x}-\bar{x}')\right]}\frac{\bar{Z}}{\left[\bar{Z} + \hat{\beta}_0^*(\bar{z}'-\bar{Z})\right]}\frac{\bar{Z}}{\left[\bar{Z} + \hat{\gamma}_0^*(\bar{z}-\bar{Z})\right]} \tag{4.8}$$

with the same MSE as given in (4.5), where $\left(\hat{C}_y, \hat{C}_x, \hat{C}_z\right)$ and $\left(\gamma_{yx,}\gamma_{yz,}\gamma_{xz,}\right)$ are sample coefficients of variations of (*y*,*x*,*z*) and sample correlation coefficients between (y and x), (y and *z*), (*x* and *z*).

From (3.8) and (4.5) we have

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}\right) - \mathrm{MSE}\left(\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0^*,\hat{\beta}_0^*,\hat{\gamma}_0^*)}\right) = \lambda S_y^2 \frac{\left(\rho_{yx}\rho_{xz} - \rho_{yz}\right)^2}{\left(1-\rho_{xz}^2\right)} > 0 \text{ provided}$$

$$\rho_{yx}, \rho_{xz} \neq \rho_{yz} \tag{4.9}$$

Thus, finally we conclude from the above studies that the estimator $\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0^*,\hat{\beta}_0^*,\hat{\gamma}_0^*)}$ is more efficient than the estimator $\bar{y}, \hat{\bar{Y}}_{Rd}, \hat{\bar{Y}}_{Rd}^{(c)}, \hat{\bar{Y}}_{Rd}^{(k)}, \hat{\bar{Y}}_{Rd}^{(u)}, \hat{\bar{Y}}_{Rd}^{(\hat{A}_0)}$ and $\hat{\bar{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}$.

Empirical Study

To see the relative performance of various estimator discussed in the paper we considered three natural population data used earlier by others. These populations are described below.

Population I (Sukhatme and Chand (1977))

N=120
y= bushels of apples harvested in 1964;
x= apples trees of bearing age in 1964;
z= bushels of apples harvested in 1959;

$\bar{Y}$ =2934.58, C_y^2 = 4.02004, C_x^2=2.5528, C_z^2=2.0379

ρ_{yx}=0.93, ρ_{yz}=0.84, ρ_{xz}=0.77

Population II (Srivastava (1971))

N=50
y= yield per plant
x= height of the plant
z= base diameter

$\bar{Y}$ =5.69 gm, C_y^2 = 0.0568, C_x^2=0.00846, C_z^2=0.01269

$$\rho_{yx}=0.7418,\ \rho_{yz}=0.5677,\ \rho_{xz}=0.2063$$

Population III (Murthy (1967))

N=34

y= area under wheat in 1964;
x= area under wheat in 1963;
z= cultivated area in 1961;

$$\overline{Y}=199.44 \text{ acres}, C_y^2=0.56728,\ C_x^2=0.51912,\ C_z^2=0.35265$$

$$\rho_{yx}=0.9801,\ \rho_{yz}=0.90430,\ \rho_{xz}=0.9097$$

The percent relative efficiencies of different estimator with respect to sample mean $\overline{y}$ are presented in Table 5.1 below.

Table 5.1. Relative efficiencies of different estimators

	Population		
	I	**II**	**III**
Estimator	n=20, n′=50	n=12, n′=20	n=7, n′=10
$\overline{y}$	100.00	100.00	100.00
$\hat{\overline{Y}}_{Rd}$	256.50	156.91	128.69
$\hat{\overline{Y}}_{Rd}^{(c)}$	509.58	730.78	181.60
$\hat{\overline{Y}}_{Rd}^{(k)}$	483.92	773.43	138.11
$\hat{\overline{Y}}_{Rd}^{(u)}$	515.31	775.18	160.43
$\hat{\overline{Y}}_{Rd}^{(\hat{A}_0)}$	520.07	778.28	193.66
$\hat{\overline{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}$	556.49	779.55	222.36
$\hat{\overline{Y}}_{Rd}^{(\hat{\alpha}^*_0,\hat{\beta}^*_0,\hat{\gamma}^*_0)}$	655.50	781.71	281.53

Table 5.1 exhibits that the performances of the proposed estimators $\hat{\overline{Y}}_{Rd}^{(\hat{A}_0)}$, $\hat{\overline{Y}}_{Rd}^{(\hat{\alpha}_0,\hat{\beta}_0)}$ and $\hat{\overline{Y}}_{Rd}^{(\hat{\alpha}^*_0,\hat{\beta}^*_0,\hat{\gamma}^*_0)}$ are better than those considered by Chand (1975), Kiregyera (1980), Upadhyaya et al (1990) and the usual estimators $\overline{y}$ and $\overline{Y}_{Rd}$. The gain in efficiency of $\hat{\overline{Y}}_{Rd}^{(\hat{\alpha}^*_0,\hat{\beta}^*_0,\hat{\gamma}^*_0)}$ is considerably more than other estimators in populations I and III while it is marginal in case of population II.

REFERENCES

CHAND,L.(1975): Some ratio-type estimators based on two or more auxiliary variables. Ph. D. Dissertation, *Iowa State University, Ames, Iowa*.

COCHRAN, W.G. (1977): Sampling Techniques. *3rd edition, John Wiley and Sons, New York.*

KIREGYERA, B.(1980): A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika,* 27,217-223.

MURTHY, M.N. (1967): Sampling Theory and Methods. *Statistical Publishing Society, Calcutta, India.*

REDDY, V.N. (1974): On a transformed ratio method of estimation. *Sankhya,* C, 36,59-70.

SAHAI,A. AND SAHAI,A. (1985): On efficient use of auxiliary information. *Jour. Statist. Plann. Inf.,* 12,203-212.

SAHOO,L.N. AND SWAIN, A.K.P.C. (1983): Chain ratio estimator. *Jour. Ind. Soc. Agril. Statist.* 35,3,70-79.

SRIVASTAVA,S.K.(1971): A generalized estimator for the mean of a finite population using multiauxiliary information. *Jour. Amer. Statist. Assoc*., 66, 404-407.

SRIVASTAVA,S.RANI, SRIVASTAVA,S.R.AND KHARE, B.B.(1989): Chain ratio-type estimator for ratio of two population means using auxiliary characters. *Commun. Statist. Theory-Method*, 18,3917-3926.

SRIVASTAVA,S.RANI, SRIVASTAVA,S.R.AND KHARE, B.B.(1990): A generalized chain ratio estimator for population mean of a finite population . *Jour. Ind. Soc. Agril. Statist.*, 42,108-117.

SRIVENKATARAMANA, T. AND TRACY, D.S. (1980): An alternative to ratio method in sample survey. *Ann. Inst. Statist. Math.,* 32,111-120.

SRIVENKATARAMANA, T. AND TRACY, D.S. (1981): Extending product methods of estimation to positive correlation case in surveys. *Aust. Jour. Statist.,* 23,95-100.

SRIVENKATARAMANA, T. AND TRACY, D.S. (1989): To phase sampling for selection with probability proportional to size in sample surveys. *Biometrika,* 76,818-821.

SUKHATME,B.V. AND CHAND,L. (1977): Multivariative ratio-type estimators. *Proceedings, Social Statistics Section, Amer. Statist. Assoc.,* 927-931.

UPADHYAYA, L.N., KUSHWAHA, K.S., SINGH,H.P.(1990): A modified chain ratio-type estimator in two phase sampling using multiauxiliary information. *Metron,* 48,381-393.

WALSH, J.E. (1970): Generalization of ratio estimate for population total. *Sankhya,* A, 32,99-106.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 1067–1079

ESTIMATION OF POPULATION RATIO IN TWO PHASE SAMPLING

G. N. Singh[1]

ABSTRACT

In this work, a two parameters family of estimators for estimating the ratio of finite population means of two different characteristics is developed utilizing the information on two auxiliary characteristics in two-phase (double) sampling. The optimum property of the suggested estimator has been studied. Theoretical and empirical studies have been made to demonstrate the efficiency of the proposed estimator with respect to the estimators, which utilize the information on one and two auxiliary variables.

***Key Words*:** Family of estimators, auxiliary information, chaining, double sampling, bias, mean square error.

1. Introduction

In sample surveys, the use of the multivariate auxiliary information in estimating the population mean of a study character has been widely made in the form of the knowledge on the population mean and on some other important population parameters. However, in many situations, the population means of all the auxiliary variables are either not known or partially known. In such situations, the usual two-phase (double) sampling procedure is used. The estimation of population mean of a study variable under the partial knowledge of the auxiliary variables means have been considered by various authors including Chand (1975), Kiregyera (1980, 84), Mukerjee et-al (1987), Srivastava et-al (1990), Singh and Singh (1991), Sahoo and Sahoo (1993), Sahoo et-al (1993), Ahmed et-al (1994), Singh et-al (1994), Singh and Upadhyaya (1995), Upadhyaya and Singh (2001) and Singh (2001). However, in many practical situations, the problem of estimating the ratio (product) of two means of two different characteristics in a finite population using information on single (or more) auxiliary variables deserves special attention, for instance, see Singh (1965, 67), Rao and Pereira

[1] Department of Applied Mathematics, Indian School of Mines, Dhanbad-826 004, INDIA. E-mail: gnsingh_ism@yahoo.com

(1968), Shah and Shah (1978), Tripathi (1980), Das (1982), Chaturvedi and Tripathi (1983), Ray and Singh (1985), Singh (1986a, 86b, 88), Srivastava et-al (1989), Singh et-al (1994a, 94b), Prasad et-al (1996) and Singh et-al (2000). In this work, a two parameters family of estimators for estimating the ratio of two means of two different characteristics has been proposed. The information on two auxiliary variables has been utilized under two-phase (double) sampling scheme. Optimum property of the proposed estimator has been discussed. Theoretical and empirical comparison of the suggested estimator has been made with the several other existing estimators.

2. Preliminary Estimators

Consider a finite population $U = (u_1, u_2, -, -, -, u_N)$ of size N in which (y_0, y_1) are the variables under study and (y_2, y_3) are the auxiliary variables such that y_2, the first auxiliary character is having good correlation with the study variables and y_3 the second auxiliary character is remotely correlated (in compare to y_2) with study variables but having good correlation with y_2. Let $\overline{Y}_i$ (i = 0, 1, 2, 3) be the population means of the respective varieties. If $\overline{Y}_1 \neq 0$, the estimation of the ratio $R = \frac{\overline{Y}_0}{\overline{Y}_1}$ is possible by using the sampling strategy [D, $\hat{R}$], where D stands for the sampling design, simple random sampling without replacement (SRSWOR) and $\hat{R}$ is the conventional estimator of R given by

$$\hat{R} = \frac{\overline{y}_0}{\overline{y}_1}; \ \overline{y}_1 \neq 0 \tag{2.1}$$

where, $\overline{y}_i$ (i = 0, 1) are the sample means of the characteristics y_i (i = 0, 1) respectively based on a sample of size n taken under D. When the information on all the units for an auxiliary characteristic y_2 is available, Singh (1965) suggested the ratio and product strategies for estimating the R
as [D, $\hat{R}_1$] and [D, $\hat{R}_2$], where

$$\hat{R}_1 = \hat{R}\frac{\overline{Y}_2}{\overline{y}_2}; \ \overline{y}_2 \neq 0 \tag{2.2}$$

and

$$\hat{R}_2 = \hat{R}\frac{\overline{y}_2}{\overline{Y}_2}; \ \overline{Y}_2 \neq 0 \tag{2.3}$$

where $\bar{y}_2$ is the sample mean of the auxiliary character y_2 based on a sample of size n under D. The strategies [D, $\hat{R}_1$] and [D, $\hat{R}_2$] are preferable over the strategy [D, $\hat{R}$] if

$$K > \frac{1}{2} \quad \text{and} \quad K < -\frac{1}{2} \tag{2.4}$$

where $K = K_{02} - K_{12}$; $K_{i2} = \rho_{i2}\frac{C_i}{C_2}$, $(i = 0,1)$, $C_t (t = 0,1,2,3)$ denote the coefficient of variation of the variate y_t (t = 0, 1, 2, 3) and $\rho_{ij} (i \neq j = 0,1,2,3)$ stands for the correlation coefficient between y_i and y_j.

In case $\bar{Y}_2$ is not known, we shall adopt the two-phase (double) sampling design with equal probability of selection and without replacement denoted as D* where the ultimate (second-phase) sample is a sub-sample of the preliminary large (first-phase) sample. Let n' and n $(n' > n)$ respectively be the sizes of the preliminary and ultimate samples. Thus the obvious double sampling strategies for estimating the R will be [D*, $\hat{R}_{1d}$] and [D*, $\hat{R}_{2d}$] where

$$\hat{R}_{1d} = \hat{R}\frac{\bar{y}_2'}{\bar{y}_2} \tag{2.5}$$

and

$$\hat{R}_{2d} = \hat{R}\frac{\bar{y}_2}{\bar{y}_2'}; \ \bar{y}_2' \neq 0 \tag{2.6}$$

where $\bar{y}_2'$ is the sample mean of the auxiliary character y_2 based on the preliminary sample of size n' drawn under D*. The strategies [D*, $\hat{R}_{1d}$] and [D*, $\hat{R}_{2d}$] respectively are preferable over [D, $\hat{R}$] under the conditions shown in (2.4).

When the information on another auxiliary character y_3 is known for each unit in the population, Chand (1975) gave a technique for chaining the estimators in the first phase sample under the sampling scheme [D*]. Using the Chand (1975) technique Singh (1989) and Srivastava et-al (1989) considered the strategies [D*, $\hat{R}_{3d}$] and [D*, $\hat{R}_{4d}$] for estimating R, where

$$\hat{R}_{3d} = \hat{R}\left(\frac{\bar{y}_2'}{\bar{y}_2}\right)\left(\frac{\bar{Y}_3}{\bar{y}_3'}\right); \ \bar{y}_3' \neq 0 \tag{2.7}$$

and

$$\hat{R}_{4d} = \hat{R}\left(\frac{\bar{y}_2}{\bar{y}_2'}\right)\left(\frac{\bar{y}_3'}{\bar{Y}_3}\right);\ \bar{Y}_3 \neq 0 \tag{2.8}$$

where $\bar{y}_3'$ is the sample mean of y_3 based on the preliminary sample of size n' drawn under D*.

Using the large sample approximations, it is easy to get the bias B(.) and mean square error (MSE), M(.) of order $o(n^{-1})$ of these estimators, which are as follows:

$$B(\hat{R}) = Rf_1\left(C_1^2 - \rho_{01}C_0C_1\right) \tag{2.9}$$

$$B(\hat{R}_1) = B(\hat{R}) + Rf_1(1-K)C_2^2 \tag{2.10}$$

$$B(\hat{R}_2) = B(\hat{R}) + Rf_1KC_2^2 \tag{2.11}$$

$$B(\hat{R}_{1d}) = B(\hat{R}) + Rf_3(\rho_{12}C_1C_2 - \rho_{02}C_0C_2) \tag{2.12}$$

$$B(\hat{R}_{2d}) = B(\hat{R}) + Rf_3(\rho_{02}C_0C_2 - \rho_{12}C_1C_2) \tag{2.13}$$

$$B(\hat{R}_{3d}) = B(\hat{R}) + R\left[f_2(C_3^2 + \rho_{13}C_1C_3 - \rho_{03}C_0C_3) + f_3(C_2^2 + \rho_{12}C_1C_2 - \rho_{02}C_0C_2\right] \tag{2.14}$$

$$B(\hat{R}_{4d}) = B(\hat{R}) + R\left[f_2(\rho_{03}C_0C_3 - \rho_{13}C_1C_3) + f_3(\rho_{02}C_0C_2 - \rho_{12}C_1C_2)\right] \tag{2.15}$$

$$M(\hat{R}) = R^2f_1(C_0^2 + C_1^2 - 2\rho_{01}C_0C_1) \tag{2.16}$$

$$M(\hat{R}_1) = M(\hat{R}) + R^2f_1C_2^2(1-2K) \tag{2.17}$$

$$M(\hat{R}_2) = M(\hat{R}) + R^2f_1C_2^2(1+2K) \tag{2.18}$$

$$M(\hat{R}_{1d}) = M(\hat{R}) + R^2f_3(C_2^2 + 2\rho_{12}C_1C_2 - 2\rho_{02}C_0C_2) \tag{2.19}$$

$$M(\hat{R}_{2d}) = M(\hat{R}) + R^2f_3(C_2^2 + 2\rho_{02}C_0C_2 - 2\rho_{12}C_1C_2) \tag{2.20}$$

$$M(\hat{R}_{3d}) = M(\hat{R}) + R^2\left[f_3(C_2^2 + 2\rho_{12}C_1C_2 - 2\rho_{02}C_0C_2) + f_2(C_3^2 + 2\rho_{13}C_1C_3 - 2\rho_{03}C_0C_3)\right] \tag{2.21}$$

$$M(\hat{R}_{4d}) = M(\hat{R}) + R^2\left[f_3(C_2^2 + 2\rho_{02}C_0C_2 - 2\rho_{12}C_1C_2) + f_2(C_3^2 + 2\rho_{03}C_0C_3 - 2\rho_{13}C_1C_3)\right] \quad (2.22)$$

where $f_1 = \left(\frac{1}{n} - \frac{1}{N}\right)$, $f_2 = \left(\frac{1}{n'} - \frac{1}{N}\right)$ and $f_3 = f_1 - f_2$.

3. The Proposed Family of Strategy

Assuming that $\overline{Y}_3$ is known, we define the sampling strategy $[D^*, \hat{R}(\alpha, \beta)]$, where

$$\hat{R}(\alpha, \beta) = \hat{R}h(\alpha)\frac{\Psi[\varphi_1(\beta)]}{\Psi[\varphi_2(\beta)]} \quad (3.1)$$

with $h(\alpha) = \left(\frac{\bar{y}_2'}{\bar{y}_2}\right)^{\alpha}$,

$$\psi[\varphi_i(\beta)] = \varphi_i(\beta) + [1 - \varphi_i(\beta)]\frac{\overline{Y}_3}{\bar{y}_3'};\ \ (i = 1, 2),$$

$$\varphi_1(\beta) = f'B/(A + f'B + C),\ \varphi_2(\beta) = C/(A + f'B + C),\ \ f' = \frac{n'}{N},$$

$$A = (\beta - 1)(\beta - 2),\ B = (\beta - 1)(\beta - 4),\ C = (\beta - 2)(\beta - 3)(\beta - 4),$$

where α is suitably chosen scalar and β (> 0) is an unknown constant. In what follows, we use the notation φ_i for $\varphi_i(\beta)$; (i = 1, 2).

Remark 1: The suggested estimator $\hat{R}(\alpha, \beta)$ describes a two-parameter family of estimators for R such that

$$\text{(i)}\ \left[D^*, \hat{R}(1,4)\right] = \left[D^*, \hat{R}_{1d}\right] \quad (3.2)$$

$$\text{(ii)}\ \left[D^*, \hat{R}(-1,4)\right] = \left[D^*, \hat{R}_{2d}\right] \quad (3.3)$$

$$\text{(iii)}\ \left[D^*, \hat{R}(1,1)\right] = \left[D^*, \hat{R}_{3d}\right] \quad (3.4)$$

$$\text{(iv)}\ \left[D^*, \hat{R}(-1,2)\right] = \left[D^*, \hat{R}_{4d}\right] \quad (3.5)$$

$$\text{(v)}\ \left[D^*, \hat{R}(1,3)\right] = \left[D^*, \hat{R}_{5d}\right] \tag{3.6}$$

$$\text{(vi)}\quad \left[D^*, \hat{R}(-1,3)\right] = \left[D^*, \hat{R}_{6d}\right] \tag{3.7}$$

where

$$\hat{R}_{5d} = \hat{R}\left(\frac{\bar{y}_2'}{\bar{Y}_2}\right)\left[(1+a) - a\frac{\bar{y}_3'}{\bar{Y}_3}\right];\quad a = n'(N-n')^{-1} \tag{3.8}$$

$$\text{and}\quad \hat{R}_{6d} = \hat{R}\left(\frac{\bar{Y}_2}{\bar{y}_2'}\right)\left[(1+a) - a\frac{\bar{y}_3'}{\bar{Y}_3}\right] \tag{3.9}$$

are the dual to ratio type estimators proposed by Srivenkataramana (1980). Thus, the strategy $[D^*, \hat{R}(\alpha,\beta)]$ may be viewed as a generalization. The bias and mean square error of $\hat{R}_{5d}$ and $\hat{R}_{6d}$ up-to $o(n^{-1})$ can be derived as

$$B(\hat{R}_{5d}) = B(\hat{R}) + R[f_3(C_2^2 + \rho_{12}C_1C_2 - \rho_{02}C_0C_2) - af_2(\rho_{03}C_0C_3 - \rho_{13}C_1C_3)] \tag{3.10}$$

$$B(\hat{R}_{6d}) = B(\hat{R}) + R[f_3(\rho_{02}C_0C_2 - \rho_{12}C_1C_2) - af_2(\rho_{03}C_0C_3 - \rho_{13}C_1C_3)] \tag{3.11}$$

$$M(\hat{R}_{5d}) = M(\hat{R}) + R^2\left[f_3(C_2^2 + 2\rho_{12}C_1C_2 - 2\rho_{02}C_0C_2) + f_2(a^2C_3^2 - 2a\rho_{03}C_0C_3 + 2a\rho_{13}C_1C_3\right] \tag{3.12}$$

$$M(\hat{R}_{6d}) = M(\hat{R}) + R^2\left[f_3(C_2^2 - 2\rho_{12}C_1C_2 + 2\rho_{02}C_0C_2) + f_2(a^2C_3^2 - 2a\rho_{03}C_0C_3 + 2a\rho_{13}C_1C_3\right] \tag{3.13}$$

4. Properties of the Proposed Strategy

To derive the bias B [.] and mean square error (MSE) M [.] of the proposed strategy, we consider the following large sample approximations:

$\bar{y}_0 = \bar{Y}_0(1+e_0), \bar{y}_1 = \bar{Y}_1(1+e_1), \bar{y}_2 = \bar{Y}_2(1+e_2),\ \bar{y}_2' = \bar{Y}_2(1+e_3), and\ \bar{y}_3' = \bar{Y}_3(1+e_4)$ such that $E(e_i) = 0$ for i = 0, 1, 2, 3, 4, we have the following results:

Theorem 1: The bias and MSE of the proposed strategy to the terms of order $o(n^{-1})$ are given by

$$B\left[\hat{R}(\alpha,\beta)\right] = B(\hat{R}) + R\left[\frac{f_1\alpha(\alpha+1)}{2}C_2^2 + \frac{f_2\alpha(\alpha-1)C_2^2}{2} - f_2\alpha^2C_2^2 + f_3\alpha(\rho_{12}C_1C_2\right.$$

$$\left. - \rho_{02}C_0C_2) + \varphi f_2(\rho_{03}C_0C_3 - \rho_{13}C_1C_3) - \varphi\varphi_2 f_2C_3^2\right] \tag{4.1}$$

$$M\left[\hat{R}(\alpha,\beta)\right] = M(\hat{R}) + R^2\left[f_2(\varphi^2 C_3^2 + 2\varphi\,\rho_{03} C_0 C_3 - 2\varphi\,\rho_{13} C_1 C_3) + f_3(\alpha^2 C_2^2 + 2\alpha\rho_{12} C_1 C_2 - 2\alpha\,\rho_{02}\, C_0 C_2)\right] \quad (4.2)$$

where $\varphi = \varphi_1 - \varphi_2$.

In order to evaluate (4.1) and (4.2) we need to assume $|\varphi_2\, e_4| < 1$. Since $\varphi_2 = C / (A + f' B + C)$, for any choice of $\beta\,(>0)$, $|\varphi_2| < 1$. So if $|e_4| < 1$, $|\varphi_2\, e_4| < 1$ is a valid assumption.

Corollary 1: The bias and MSE of the strategies given in (3.2) – (3.7) can be obtained by taking the suitable choices of α and β in (4.1) and (4.2) respectively.

Theorem 2: The strategy $\left[D^*, \hat{R}(\alpha,\beta)\right]$, its bias and MSE are asymptotically convergent to the strategy $\left[D^*, \hat{R}(\alpha,1)\right]$, its bias and MSE respectively for large β.

where $\hat{R}(\alpha,1) = \hat{R}h(\alpha)\dfrac{\overline{Y}_3}{\overline{y}_3'}$

Proof: Since $\beta\,(>0)$, dividing $\psi[\varphi_1]$ and $\psi[\varphi_2]$ in (3.1) by β^3 and taking limit as $\beta \to \infty$, we have $\psi[\varphi_1] = \dfrac{\overline{Y}_3}{\overline{y}_3'}$ and $\psi[\varphi_2] = 1$. Thus,

$$\lim_{\beta\to\infty} \hat{R}(\alpha,\beta) = \hat{R}(\alpha,1) \quad (4.3)$$

$$\lim_{\beta\to\infty} B\left[\hat{R}(\alpha,\beta)\right] = B\left[\hat{R}(\alpha,1)\right] \quad (4.4)$$

$$\lim_{\beta\to\infty} M\left[\hat{R}(\alpha,\beta)\right] = M\left[\hat{R}(\alpha,1)\right] \quad (4.5)$$

This completes the proof.

The proposed two parameters family of strategy, therefore, converges to one parameter family of strategy even if one chooses arbitrarily a large value of the unknown parameter β. This property distinguishes the proposed strategy with other two parameters strategies, which generally fail to exhibit such convergence.

Theorem 3: The optimum Choices of α and β which minimize $M[\hat{R}(\alpha,\beta)]$ are

$$\alpha = K \quad (4.6)$$

and the real and positive roots of the equation

$$\varphi = \mathrm{K}^{*} \tag{4.7}$$

where $\mathrm{K}^{*} = \rho_{13}\frac{C_1}{C_3} - \rho_{03}\frac{C_0}{C_3}$

and minimum MSE is given by

$$\mathrm{M}[\hat{\mathrm{R}}(\alpha,\beta)]_0 = \mathrm{M}(\hat{\mathrm{R}}) - \mathrm{R}^2[\mathrm{f}_2(\rho_{03}\mathrm{C}_0 - \rho_{13}\mathrm{C}_1)^2 + \mathrm{f}_3(\rho_{03}\mathrm{C}_0 - \rho_{12}\mathrm{C}_1)^2] \tag{4.8}$$

Remark 1: A close look of the equation (4.7) reveals that it is a cubic equation in β. Therefore, for any given K* one will get three optimum values of β for which $\mathrm{M}[\hat{\mathrm{R}}(\alpha,\beta)]$ attains same minimum value. The possibility of getting negative or imaginary roots cannot be ruled out. However, it could be seen that for any choice of f′ and K* there exists at least one positive real root of the equation ensuring that $\mathrm{M}[\hat{\mathrm{R}}(\alpha,\beta)]$ attains its minimum within the parametric space (0, ∞).

Remark 2: Since there may exist at the most three optimum values of β, a criterion for selecting suitable value of optimum β may be set as follows:

"Out of all the possible values of optimum β, select that β which makes $|B[\hat{R}(\alpha,\beta)]|$ smallest".

Remark 3: It is interesting to note that (4.6) and (4.7) need the knowledge of $\rho_{i2}\frac{C_i}{C_2}(i=0,1)$ and $\rho_{i3}\frac{C_i}{C_3}(i=0,1)$. It is advocated in the literature see Reddy (1978) that these quantities do not fluctuate much over time and over repeated surveys and hence their values may be assessed accurately from pilot surveys or past experience.

5. Efficiency Comparisons

Comparing the expression (4.8) with (2.19), (2.20), (2.21), (2.22), (3.12) and (3.13), we have the following:

Theorem 4: $\mathrm{M}[\hat{\mathrm{R}}(\alpha,\beta)]_o$ is always smaller than the MSE's of $\hat{\mathrm{R}}_{1\mathrm{d}}$, $\hat{\mathrm{R}}_{2\mathrm{d}}$, $\hat{\mathrm{R}}_{3\mathrm{d}}$, $\hat{\mathrm{R}}_{4\mathrm{d}}$, $\hat{\mathrm{R}}_{5\mathrm{d}}$ and $\hat{\mathrm{R}}_{6\mathrm{d}}$.

In order to decide the range of the parametric space where the suggested estimator is better than the estimators $\hat{\mathrm{R}}_{1\mathrm{d}}$, $\hat{\mathrm{R}}_{2\mathrm{d}}$, $\hat{\mathrm{R}}_{3\mathrm{d}}$, $\hat{\mathrm{R}}_{4\mathrm{d}}$, $\hat{\mathrm{R}}_{5\mathrm{d}}$ and $\hat{\mathrm{R}}_{6\mathrm{d}}$, we compare the MSE of $\hat{\mathrm{R}}(\alpha,\beta)$ given in (4.2) with the MSE's of $\hat{\mathrm{R}}_{1\mathrm{d}}$, $\hat{\mathrm{R}}_{2\mathrm{d}}$, $\hat{\mathrm{R}}_{3\mathrm{d}}$,

$\hat{R}_{4d}$, $\hat{R}_{5d}$ and $\hat{R}_{6d}$ given in the expressions (2.19), (2.20), (2.21), (2.22), (3.12) and (3.13) respectively. We have the following results:

(i) $\hat{R}(\alpha,\beta)$ is better than $\hat{R}_{1d}$ if

$$\varphi < 2K^{*} \text{ and } \alpha < 2K - 1 \tag{5.1}$$

(ii) $\hat{R}(\alpha,\beta)$ is preferable over $\hat{R}_{2d}$ if

$$\varphi < 2K^{*} \text{ and } \alpha < 2K + 1 \tag{5.2}$$

(iii) $\hat{R}(\alpha,\beta)$ is superior to $\hat{R}_{3d}$ if

$$\varphi < 2K^{*} + 1 \text{ and } \alpha < 2K - 1 \tag{5.3}$$

(iv) $\hat{R}(\alpha,\beta)$ *dominates* $\hat{R}_{4d}$ *if*

$$\varphi < 2K^{*} - 1 \text{ and } \alpha < 2K + 1 \tag{5.4}$$

(v) $\hat{R}(\alpha,\beta)$ *performs better than* $\hat{R}_{5d}$ *if*

$$\varphi < 2K^{*} + a \text{ and } \alpha < 2K - 1 \tag{5.5}$$

(vi) $\hat{R}(\alpha,\beta)$ *dominates* $\hat{R}_{5d}$ *if*

$$\varphi < 2K^{*} + a \text{ and } \alpha < 2K + 1 \tag{5.6}$$

6. Numerical Illustration

The various results obtained in previous sections are now examined with the help of data earlier considered by Srivastava et-al (1988). In this data fifty-five children of 5 years age belonging to Varanasi district of India were considered to estimate the index for malnutrition, i.e., weight/mid arm circumference as $(\overline{Y}_o / \overline{Y}_1)$. The auxiliary characteristics chosen for this purpose is the chest circumference (y_2) and skull circumference (y_3). For this population required values are.

$\overline{Y}_o = 17.08$, $\overline{Y}_1 = 16.92$, $\overline{Y}_3 = 50.44$, $C_0 = 0.1269$, $C_1 = 0.0831$

$C_2 = 0.0520$, $C_3 = 0.0265$, $\rho_{01} = 0.54$, $\rho_{02} = 0.84$, $\rho_{12} = 0.78$, $\rho_{13} = 0.50$,

$\rho_{03} = 0.51$, $\rho_{23} = 0.65$, $N = 55$, $n' = 22$, $n = 15$.

Table 1 shows the optimum values of α and β with corresponding bias and percent relative efficiencies of $\hat{R}(\alpha,\beta)$ with respect to the various estimators in two-phase sampling scheme D*.

Table 1.

α_{opt}	β_{opt}	B($\hat{R}(\alpha\beta)$)	$\hat{R}(\alpha\beta)$	$\hat{R}_{1d}$	$\hat{R}_{2d}$	$\hat{R}_{3d}$	$\hat{R}_{4d}$	$\hat{R}_{5d}$	$\hat{R}_{6d}$
0.8034	1.342885*	0.000066	100	101.82	145.45	181.82	200.00	105.45	145.45
	2.215920	0.000207	100						
	15.391195	0.000067	100						

Table 2 shows the ranges of α and φ for which $\hat{R}(\alpha,\beta)$ is preferable over the various estimators in two-phase sampling.

Table 2.

Estimators	$\hat{R}_{1d}$	$\hat{R}_{2d}$	$\hat{R}_{3d}$	$\hat{R}_{4d}$	$\hat{R}_{5d}$	$\hat{R}_{6d}$
φ is less than	-1.7486	-1.7486	-0.7486	-2.7486	-1.0819	-1.0819
α is less than	0.6068	2.6068	0.6068	2.6068	0.6068	2.6068

From the table 1, it is clear that β = 1.342885 is the best choice according to the criterion given in Remark 2. In general, the use of additional auxiliary characteristics y_3 makes the estimator more efficient then the estimators, which do not require such knowledge.

Table 2, depicts the ranges of φ and α in which $M\left[\hat{R}(\alpha,\beta)\right]$ is smaller than the MSE's of other estimators under consideration. Thus, any choice of φ and α in these ranges will make the estimator $\hat{R}(\alpha,\beta)$ superior to the estimators considered in the table 2.

7. Conclusions

From the above results, it can be concluded that the proposed strategy has some specific properties, which makes it preferable over other sampling strategies utilizing one or two auxiliary characteristics. Although the guessing of the unknown parameters depend upon the certain population parameters, the loss in efficiency is negligible if these are estimated with the help of sample values.

REFERENCES

AHMED, M. S., KHAN, S. U. and TRIPATHI, T. P. (1994): Two general class of chain ratio and product estimators for a finite population mean based on two phase sampling and multivariate information. J. Statistical Studies, 14, 86-99.

CHAND, L. (1975): Some ratio type estimators based on two or more auxiliary variables. Unpublished Ph. D. thesis, Iowa State University, Ames, Iowa (USA).

CHATURVEDI, D. K. and TRIPATHI, T. P. (1983): Estimation of population ratio on two occasions using multi-auxiliary information. J. Indian Statist. Assoc., 21, 113-120.

DAS, A. K. (1982): Estimation of population ratio on two occasions. Jour. Indian Soc. Agricultural Statist., 34, 1-9.

KIREGYERA, B. (1980): A chain ratio type estimators in finite population double sampling using two auxiliary variables. Metrika, 17, 217-223.

KIREGYERA, B. (1984): Regression type estimators using two auxiliary variables and the model of double sampling from finite populations. Metrika, 31, 215-226.

MUKERJEE, R., RAO, T. J. and VIJAYAN, K. (1987): Regression type estimators using multiple auxiliary information. Austral. Jour. of Statist., 29, 244-254.

PRASAD, B., SINGH, R. S. and SINGH, H. P. (1996): Some chain ratio-type estimators for ratio of two population means using two auxiliary characters in two-phase sampling. Metron, 54, 95-113.

RAO, J. N. K. and PEREIRA, N. P. (1968): On double ratio estimators. Sankhya, Ser. A, 30, 83-90.

RAY, S. K. and SINGH, R. K. (1985): Some estimators for the ratio and product of population parameters. Jour. Indian Soc. Agricultural Statist., 37, 1-10.

REDDY, V. N. (1978): A study on the use of prior knowledge on certain population parameters in estimation. Sankhya, Ser. C, 40, 29-37.

SAHOO, J. and SAHOO, L. N. (1993): A class of estimators in two phase sampling using two auxiliary variables. J. Indian Statist. Assoc. 31, 107-114.

SHAOO, J., SAHOO. L. N. and MOHANTY, S. (1993): A regression approach to estimation in two phase sampling using two auxiliary variables. Current Science, 65, 73-75.

SHAH, S. M. and SHAH, D. N. (1978): Ratio-cum-product estimators for estimating ratio (product) of two population parameters. Sankhya, Ser. C., 40, 156-166.

SINGH, M. P. (1965): On the estimation of ratio and product of the population parameters. Sankhya, Ser. B, 27, 321-328.

SINGH, M. P. (1967): Ratio-cum-product method of estimation. Metrika, 12, 34-43.

SINGH, H. P. (1986a): Estimation of ratio, product and mean using auxiliary information in sample surveys. Aligarh J. Statist., 6, 32-44.

SINGH, H. P. (1986b): A generalized class of estimators of ratio, product and mean using supplementary information on an auxiliary character in PPSWR sampling scheme. Gujarat Statist. Rev., 13, 1-30.

SINGH, H. P. (1988): On the estimation of ratio and product of two finite population means. Proc. Nat. Accd. Sci. India, Sec. A., 58, 399-402.

SINGH, V. P. (1989): Use of auxiliary information in estimation of population parameters. M.Sc. dissertation, JNKVV, Jabalpur, M. P., India.

SINGH, V. K. and SINGH, G. N. (1991): Chain type regression estimators with two auxiliary variables under double sampling schemes. Metron, 49, 279-289.

SINGH, V. K., SINGH, G. N. and SHUKLA, D. (1994): A class of chain ratio type estimators with two auxiliary variables under double sampling scheme. Sinkhya, Ser. B, 56, 209-221.

SINGH, V. K., SINGH, H. P., SINGH, H. P. and SHUKLA, D. (1994a): A general class of chain estimators for ratio and product of two means of a finite population. Commun. Statist-Theory Meth., 23, 1341-1355.

SINGH, V. K., SINGH, H. P. and SINGH, H. P. (1994b): Estimation of ratio and product of two finite population means in two phase sampling. J. Statist. Planning Inference, 41, 163-171.

SINGH, G. N. and UPADHYAYA, L. N. (1995): A class of modified chain-type estimators using two auxiliary variables in two-phase sampling. Metron, LIII, No.3-4, 117-125.

SINGH, G. N., SINGH, V. K. and UPADHYAYA, L. N. (2000): On a family of estimator for population ratio in sample surveys. Jour. Kerala Stat. Assoc., 11, 12-18.

SINGH, G. N. (2001): On the use of transformed auxiliary variable in estimation of population mean in two phase sampling. Statistics in Transition, 5, No.3, 405-416.

SRIVASTAVA, S. R., KHARE, B. B. and SRIVASTAVA, S. R. (1988): On generalized chain estimator for ratio and product of two population means using auxiliary charactes. Assam Statist. Rev., 2, 21-29.

SRIVASTAVA, S. R. SRIVASTAVA, S. R. and KHARE, B. B. (1989): Chain ratio type estimators for ratio of two population means using auxiliary characters. Commun Statist. – Theory Meth., 18, 3917-3926.

SRIVASTAVA, S. R., KHARE, B. B. and SRIVASTAVA, S. R. (1990): A generalized chain ratio estimator for mean of finite population. Jour. Indian Soc. Agricultural Statist, 42, 108-117.

SRIVENKATARAMANA, T. (1980): A dual to ratio estimator in sample surveys. Biometrika, 67, 199-204.

TRIPATHI, T. P. (1980): A general class of estimators for population ratio. Sankhya, Ser. C., 42, 63-75.

UPADHYAYA, L. N. and SINGH, G. N. (2001): Chain-type estimators using transformed auxiliary variable in two-phase sampling. A.M.S.E., 38, No.1,2, 1-9.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 887–

A GENERAL METHOD OF ESTIMATION AND ITS APPLICATION TO THE ESTIMATION OF COEFFICIENT OF VARIATION

T. P. Tripathi[1], H. P. Singh[2] and L. N. Upadhyaya[33]

ABSTRACT

A general class of estimators for estimating any specified population parameter θ_0 using apriori information on some other parameters θ_1 and θ_2 is proposed, its general properties are studied and the lower bound to the mean square error of the estimators in the class is obtained under large sample approximation. The general results are then applied to estimate the coefficient of variation of a principal character y using knowledge of the population mean and variance of an auxiliary character x. An empirical illustration is given.

Key words: Auxiliary character, mean-squared error, relative efficiency.

1. Introduction

The use of auxiliary or apriori information in obtaining 'desirable' sampling strategies has been invariably recognized by various authors in sample surveys. The use of ratio, regression and product methods of estimation are good examples in this context. Several successful attempts have been made to estimate a specified population parameter θ_0 (e.g. $\overline{Y}, \sigma_y^2$ and C_y , the population mean, variance and coefficient of variation of the study character y) using apriori information on a single parameter θ_1 (e.g. $\overline{X}, \sigma_x^2$ and C_x , the population mean, variance and coefficient of variation of the study character x) by a number of statisticians.

The problem of estimating a parameter θ_0 using knowledge of some other parameter θ_1 was dealt under a general frame work by Das and Tripathi (1980) who discussed a class of estimators defined as

[1] Indian Statistical Institute, Kolkata-700 035. India.
[2] Vikram University, Ujjain- 456 010. India.
[3] Indian School of Mines, Dhanbad-826 004. India.

$$d_1 = \frac{\left[\hat{\theta}_0 - t_1(\hat{\theta}_1 - \theta_1)\right]}{\left[\hat{\theta}_1 - t_2(\hat{\theta}_1 - \theta_1)\right]} (\theta_1)^{\alpha} \tag{1.1}$$

where t_1 and t_2 are suitably chosen statistics, α is a suitably chosen constant $\hat{\theta}_0$ and $\hat{\theta}_1$ are the estimators of θ_0 and θ_1 respectively based on any probability sampling design.

Following Srivastava (1971, 80) one may also in general define the class of estimators for θ_0 as

$$d_2 = \hat{\theta}_0 \, h(u_1) \tag{1.2}$$

$$d_3 = g(\hat{\theta}_0, \hat{\theta}_1, \theta_1) \tag{1.3}$$

where $u_1 = (\hat{\theta}_1/\theta_1)$ and h(.) and g (.) are functions which satisfy regularity conditions similar to those assumed by Srivastava (1971, 80).

In case $\hat{\theta}_0$ and $\hat{\theta}_1$ are unbiased estimators, the minimum mean square error (MSE) of the estimators in the class d_1, d_2 and d_3 is found to be

$$M_0(d_i) = (1 - \rho_{01}^2) \, V(\hat{\theta}_0), (i = 1,2,3) \tag{1.4}$$

which is the MSE of regression type estimators

$$d_0 = \hat{\theta}_0 - \hat{\beta}_{01}(\hat{\theta}_1 - \theta_1)$$

where ρ_{01} is the correlation coefficient between $\hat{\theta}_0$ and $\hat{\theta}_1$ and $\hat{\beta}_{01}$ is an estimator of β_{01}, the regression coefficient of $\hat{\theta}_0$ on $\hat{\theta}_1$.

The estimators in the class d_1, d_2, d_3 may further be improved in many survey situations, where it is possible to have apriori information on two parameters θ_1 and θ_2 such as $\overline{X}$ and C_x, $\overline{X}$ and σ_x^2 and C_y and $\overline{X}$ etc. while θ_0 may be $\overline{Y}$. In such situations Das and Tripathi (1981 C) suggested a class of estimators for θ_0 as

$$d_4 = \hat{\theta}_0 - t_1^* (\hat{\theta}_1 - \theta_1) - t_2^* (\hat{\theta}_2 - \theta_2) \tag{1.5}$$

where $\hat{\theta}_i$ (i = 0, 1, 2) is an estimator for θ_1 based on any general sampling design, t_1^* and t_2^* are suitably chosen statistics, which may be, in particular, a constant.

With the same amount of information one may define a general class of estimators for θ_0 as

$$d_g^* = g(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2) \tag{1.6}$$

In this paper we study the general properties of the class d_g^* and obtain lower bound to the mean square errors of the estimators in the class. The results are used, in particular, for estimating C_y, the coefficient of variation of study character y in case both $\overline{X}$ and σ_x^2 are known.

2. Properties of the proposed class of estimators

To study the properties of d_g^* we assume that :

(i) $t = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2)$ assume values in a closed convex subset, S, of the three dimensional real space containing the point $T = (\theta_0, \theta_1, \theta_2)$.

(ii) $g(.)$ is a function of $\hat{\theta}_0$, t such that

$$g(t)|_{t=T} = \theta_0 \tag{2.1}$$

(iii) the function g(.) is continuous and bounded in S.

(iv) the first and second order partial derivatives of g(.) exist and are continuous and bounded.

Expanding g(.) about the point T in a second order Taylor's series, we have

$$d_g^* = \theta_0 + \varepsilon_0 g_1(t) + \varepsilon_1 g_2(t) + \varepsilon_2 g_3(t)$$

$$+ \frac{1}{2}\{\varepsilon_0^2\, g_{11}^2(t^*) + 2\,\varepsilon_0\,\varepsilon_2 g_{12}(t^*) + \varepsilon_1^2\, g_{22}(t^*)$$

$$+ 2\,\varepsilon_0\,\varepsilon_2 g_{13}(t^*) + 2\,\varepsilon_1\,\varepsilon_2\, g_{23}(t^*) + \varepsilon_2^2\, g_{33}(t^*)\ \} \tag{2.2}$$

where $\varepsilon_i = (\hat{\theta}_i - \theta_i), (i = 0,1,2)$ with $E(\varepsilon_i) = 0\ \forall\ i = 0,1,2$,

$\theta_i^* = (\hat{\theta}_0^*, \hat{\theta}_1^*, \hat{\theta}_2^*, \theta_1, \theta_2), \hat{\theta}_i^* = \theta_i + \theta\varepsilon_i,\ (i = 0,1,2), 0 < \theta_i < 1$ and g_{ii} (i, j = 1, 2, 3) denote partial derivatives of the function g(.) and

g_{ij} (i , j = 1, 2, 3) its second order partial derivatives. We shall assume that

$$E(\hat{\theta}_0 - \theta_0)^{r_1}(\hat{\theta}_1 - \theta_1)^{r_2}(\hat{\theta}_2 - \theta_2)^{r_3} = \begin{cases} 0(n^{-1}), \text{for } r_1 + r_2 + r_3 = 2 \\ 0(n^{-q}), q > 1, \text{for } r_1 + r_2 + r_3 > 2 \end{cases} \tag{2.3}$$

where $r_i's\ (i = 1, 2, 3)$ are positive integers.

Taking expectation of (2.2) it is found that

$$E(d_g^*) = \theta_0 + B(\hat{\theta}_0)\, g_1(T) + B(\hat{\theta}_1)\, g_2(T) + B(\hat{\theta}_2)\, g_3(T) + 0(n^{-1}),$$

and hence the bias of d_g^* would be

$$B(d_g^*) = B(\hat{\theta}_0)\, g_1(T) + B(\hat{\theta}_1)\, g_2(T) + B(\hat{\theta}_2)\, g_3(T) + 0(n^{-1}), \tag{2.4}$$

where n is the sample size.

Noting from (2.1) that $g_1(T) = 1$, we obtain the MSE of $\hat{d}_g$ to the terms of the order $0(n^{-1})$, as

$$\begin{aligned} M(d_g^*) &= M(\hat{\theta}_0) + M(\hat{\theta}_1) g_2^2(T) + M(\hat{\theta}_2)\, g_3^2(T) + 2\, C(\hat{\theta}_0, \hat{\theta}_1)\, g_2(T) \\ &+ 2\, C(\hat{\theta}_0, \hat{\theta}_2)\, g_3(T) + 2C(\hat{\theta}_1, \hat{\theta}_2)\, g_3(T) + 2C(\hat{\theta}_1, \hat{\theta}_2)\, g_2(T)\, g_3(T) \end{aligned} \tag{2.5}$$

where $C(\hat{\theta}_i, \hat{\theta}_j) = COV(\hat{\theta}_i, \hat{\theta}_j) + B(\hat{\theta}_i)\, B(\hat{\theta}_j)$, $(i = j = 0, 1, 2)$.

Minimising the MSE of $\hat{d}_g^*$ in (2.5) with respect to $g_2(T)$ and $g_3(T)$, we obtain

$$\left.\begin{aligned} g_2(T) &= \frac{\left[C(\hat{\theta}_0, \hat{\theta}_2)\, C(\hat{\theta}_1, \hat{\theta}_2) - M(\hat{\theta}_2)\, C(\hat{\theta}_0, \hat{\theta}_1)\right]}{\left[M(\hat{\theta}_1)\, M(\hat{\theta}_2) - \left(C(\hat{\theta}_1, \hat{\theta}_2)\right)^2\right]} \\ g_3(T) &= \frac{\left[C(\hat{\theta}_0, \hat{\theta}_1)\, C(\hat{\theta}_1, \hat{\theta}_2) - M(\hat{\theta}_1)\, C(\hat{\theta}_0, \hat{\theta}_2)\right]}{\left[M(\hat{\theta}_1)\, M(\hat{\theta}_2) - \left(C(\hat{\theta}_1, \hat{\theta}_2)\right)^2\right]} \end{aligned}\right\} \tag{2.6}$$

and thus the minimum MSE of d_g^*, designated as $M_0(d_g^*)$, to terms of order 0 (n^{-1}), is given by

$$M_0(d_g^*) = M_0(\hat{\theta}_0) - M, \tag{2.7}$$

where

$$M = \frac{M(\hat{\theta}_1)\left\{C(\hat{\theta}_0, \hat{\theta}_2)\right\}^2 + M(\hat{\theta}_2)\left\{C(\hat{\theta}_0, \hat{\theta}_1)\right\}^2 - 2C(\hat{\theta}_0, \hat{\theta}_2)\, C(\hat{\theta}_1, \hat{\theta}_2)}{\left[M(\hat{\theta}_1)\, M(\hat{\theta}_2) - \left\{C(\theta_1, \theta_2\right\}^2\right]}$$

The minimum MSE $M_0(d_g^*)$ is identical to the asymptotic optimum MSE of the randomly weighted estimator d_4 in (.15). Thus, in the class obtained by the

union of classes d_4 and d_g^*, no estimator would have MSE less than $M_0(d_g^*) = M_0(d_4)$.

For illustration, in case θ_1 and θ_2 are known, we can have some interesting class of estimators which are sub class of d_g^*, such as:

$$e_1 = \sum_{i=1}^{2} w_i \left\{\hat{\theta}_0 - \lambda_i(\hat{\theta}_i - \theta_i)\right\}, \sum_{i=1}^{2} w_i = 1 \tag{2.8}$$

$$e_2 = w_0\,\hat{\theta}_0 + w_1(\hat{\theta}_1 - \theta_1) + w_2(\hat{\theta}_2 - \theta_2), \sum_{i=0}^{2} w_i = 1 \tag{2.9}$$

$$e_3 = \hat{\theta}_0 - \sum_{i=1}^{2} \lambda(\hat{\theta}_i - \theta_i), \ (i = 1, 2) \tag{2.10}$$

$$e_4 = \hat{\theta}_0 u_1^{\alpha_2}\, u_2^{\alpha_2}, u_i = \hat{\theta}_i/\theta_i, \quad (i = 1, 2) \tag{2.11}$$

$$e_5 = \left[\hat{\theta}_0 - \lambda_1(\hat{\theta}_1 - \theta_1)\right](\hat{\theta}_2/\theta_2)^{\alpha} \tag{2.12}$$

$$e_6 = \frac{\left[\hat{\theta}_0 - \lambda_1(\hat{\theta}_1 - \theta_1) - \lambda_2(\hat{\theta}_1 - \theta_2)\right]}{\left[\hat{\theta}_1 - \lambda_1^*(\hat{\theta}_1 - \theta_1) - \lambda_2^*(\hat{\theta}_2 - \theta_2)\right]}\{\theta_1\} \tag{2.13}$$

$$e_7 = \hat{\theta}_0 \frac{\left[1 + \alpha_1(u_1 - 1)\right]}{\left[1 + \alpha_2(u_2 - 1)\right]} \tag{2.14}$$

$$e_8 = \frac{\hat{\theta}_0}{\left[1 + \alpha_1(u_1 - 1) - \alpha_2(u_2 - 1)\right]} \tag{2.15}$$

$$e_9 = \hat{\theta}_0 \left[w_1 u_1 + w_2 u_2^{\alpha_2}\right], \sum_{i=1}^{2} w_i = 1 \tag{2.16}$$

etc., where $(\lambda_1, \lambda_2, \lambda_1^*, \lambda_2^*)$, $(\alpha, \alpha_1, \alpha_2)$ and (w_0, w_1, w_2) are suitably chosen constants.

REMARKS:

(i) The estimators d_i (i = 1, 2, 3) defined in (1.1) to (1.3) are the sub class of the proposed class of estimators d_g^*. The common minimum MSE of the estimators d_1, d_2 and d_3 is given by

$$M_0(d_i) = M(\hat{\theta}_0) - \frac{\{C(\theta_0, \theta_1)\}^2}{M(\hat{\theta}_1)} \quad , \quad (i = 1, 2, 3) \tag{2.17}$$

It is observed from (2.7) and (2.17) that the asymptotically optimum estimators in d_g^* would be better than all the estimators in d_1, d_2 and d_3.

(i) It is interesting to note that the above results given in (2.6) and (2.7) are very general in nature and can be used for estimating θ_0 whether population is finite or infinite. Further, they can be used to estimate:

(a) the population mean $\overline{Y}$ when ($\overline{X}$, C_x) or ($\overline{X}$, σ_x^2) is known. [see Das and Tripathi (1981 C) and Srivastava and Jhajj (1980 a)],

(b) the population variance σ^2 when ($\overline{X}$, C_x) or ($\overline{X}$, σ_x^2) is known. [see Srivastava and Jhajj (1980 a)],

(c) the population coefficient of variation C_y when ($\overline{X}$,C_x) or ($\overline{X}$,σ_x^2) is known. [see section 3],

(d) the population correlation coefficient ρ when ($\overline{X}$, C_x) or ($\overline{X}$,σ_x^2) is known. [see Srivastava and Jhajj (1986)],

(e) the population regression coefficient β (= σ_{xy}/σ_x^2) when ($\overline{X}$,C_x) or ($\overline{X}$, σ_x^2) is known,

(f) $\overline{Y}^2$ when ($\overline{X}$, C_x) or ($\overline{X}$, σ_x^2) is known,

(g) the ratio of two finite population means R (=$\overline{Y}_1/\overline{Y}_2$, $\overline{Y}_2 \neq 0$) when ($\overline{X}$, C_x) or ($\overline{X}$, σ_x^2) is known [see Upadhyaya and Singh (1985),

(h) inverse of the population mean $(1/\overline{Y},\ \overline{Y} \neq 0)$ when ($\overline{X}$,C_x) or ($\overline{X}$, σ_x^2) is known.

3. Estimators of C_y using information on mean and variance of an auxiliary character

The problem of estimating population mean $\overline{Y}$ (or total $Y = N\overline{Y}$) has been considered extensively in the sample survey literature.

The problem of estimating the variance σ_y^2 has been considered, among others by Wakimoto (1971), Singh et al (1973), Liu (1974), Das and Tripathi (1977, 78), Srivastava and Jhajj (1980), Searls and Interapanich (1990) and Singh et al (1988, 90).

In many situations the problem of estimating the coefficient of variation C_y of study character y in case of finite (or infinite) populations assumes importance and is not merely of theoretical interest devoid of any practical necessity. The

parameter C_y as a measure of variability of study character y, or as a measure of dispersion per unit mean, in the population is one of the important parameters. The problem of estimation of Cy using information on auxiliary character x has been first taken up by Das and Tripathi (1981 a, b, 1992).

The estimator $\hat{C}_y$, in general is biased for large samples. The expressions for its bias and MSE are given by

$$B(\hat{C}_y) = K\, C_y \left[C_y^2 - (1/2)\, C_{12}(y,y) - (1/8)\,(A\,\beta_2(y) - B)\right] \tag{3.1}$$

$$M(\hat{C}_y) = K\, C_y^2 \left[C_y^2 - C_{12}(y,y) + (1/4)\,(A\,\beta_2(y) - B)\right] \tag{3.2}$$

where $K = (N-n)/\{n(N-1)\}, \mu_r(y) = \sum_{i=1}^{N}(y_i - \overline{Y})^r/N, \quad (r = 2, 3,4)$

$C_{12}(y, y) = \{(N-1)/(N-2)\}\mu_3(y)/(\overline{Y}\,\sigma_y^2),\ \beta_2(y) = \mu_4/\sigma_y^4,$

$$A = \left(\frac{N-1}{N}\right)^2\left[\frac{n^2}{(n-1)^2} - \frac{2\,n\,(N-2n)}{(n-1)^2(N-2)} + \frac{(N^2+N-6nN+6n^2}{(n-1)^2(N-2)(N-3)}\right],$$

$$B = \left(\frac{N-1}{N}\right)^2\left[\frac{n\,N^2}{(n-1)(N-1)(N-2)} - \frac{3\,N\,(N-n-1)}{(n-1)(N-2)(N-3)}\right].$$

Das and Tripathi (1981 a, b) presented two classes of estimators based on SRSWOR for C_y in different situations as

$$D_1 = \frac{\left[\hat{C}_y - \lambda_1(\overline{x} - \overline{X})\right]}{\left[\overline{x} - \lambda_2(\overline{x} - \overline{X})\right]^{\alpha_1}}(\overline{X})^{\alpha_1}, \quad \text{when } \overline{X} \text{ is known,} \tag{3.3}$$

and

$$D_2 = \frac{\left[\hat{C}_y - \lambda_3(\hat{C}_x^2 - C_x^2)\right]}{\left[\hat{C}_x^2 - \lambda_4(\hat{C}_x^2 - C_x^2)\right]^{\alpha_2}}(C_x^2)^{\alpha_2}, \quad \text{when } C_x^2 \text{ is known,} \tag{3.4}$$

where λ_i (i = 1,2,3,4) and α_i = (i = 1,2,) are suitably chosen constants

and $\hat{C}_x^2 = \left(\frac{N-1}{N}\right)^{1/2} s_x/\overline{x}.$

The minimum MSE's of D_1 and D_2 are respectively given by

$$M_0(D_1) = M(\hat{C}_y) - K(C_y/C_x)^2 A_2^2 \tag{3.5}$$

$$M_0(D_1) = M(\hat{C}_y) - \frac{KC_y^2(A_1 - 2A_2)^2}{\left[4\{C_x^2 - C_{12}(x,x)\}\right]} \tag{3.6}$$

where

$$\mu_{ab} = \sum_{i=1}^{N}(y_i - \overline{Y})^a\ (x_i - \overline{X})^b/N, \beta_2(x) = \mu_4(x)/\sigma_x^4,$$

$$C_{22}(y,x) = \mu_{22}(y,x)/(\sigma_y^2\ \sigma_x^2),$$

$$C_{12}(y,x) = \{(N-1)/(N-2)\}\mu_3(x)/(\overline{X}\ \sigma_x^2),$$

$$C = \left(\frac{N-1}{N}\right)^2 \frac{2N(N-n-1)}{(n-1)(N-2)(N-3)},$$

$$g(y,x) = [A\ C_{22}(y,x) - B - C(1-\rho^2)],$$

$$A_1 = [(1/2)\ g(y,x) - C_{12}(y,x)],$$

$$A_2 = [(1/2)\ C_{21}(y,x) - \rho\ C_y\ C_x],$$

$$B_1 = [A\ \beta_2(x) - B],$$

$$C_{21}(y,x) = \{(N-1)/(N-2)\}\ \mu_{21}(y,x)/(\overline{X}\ \sigma_y^2),$$

$$C_{12}(y,x) = \{(N-1)/(N-2)\ \mu_{12}(y,x)\ (\overline{Y}\ \sigma_x^2).$$

In many situations knowledge on both $\overline{X}$ and σ_y^2 (or equivalently, $\overline{X}$ and C_x) may be available which may be used to provide estimator better than $\hat{C}_y$. In such a situation we define a class of estimators for C_y, based on the SRSWOR, as

$$d_g = g(\hat{C}_y, \overline{x}, s_x^2, \overline{X}, \sigma_x^2) \tag{3.7}$$

The results for d_g may be derived from (2.6) and (2.7) taking $\hat{\theta}_1$ and $\hat{\theta}_2$ as unbiased estimators of θ_1 and θ_2 respectively.

Thus, in case of sampling scheme SRSWOR, we get

$$\left.\begin{aligned} g_2(T^*) &= \frac{C_y}{\overline{X}}\,\frac{\left[C_{12}(x,x)\,A_1 - A_2B_1\right]}{\left[B_1C_x^2 - C_{12}^2(x,x)\right]} \\ g_3(T^*) &= \frac{C_y}{\sigma_x^2}\,\frac{\left[C_{12}(x,x)\,A_2 - C_x^2A_1\right]}{\left[B_1C_x^2 - C_{12}^2(x,x)\right]} \end{aligned}\right] \tag{3.8}$$

and

$$M_0(d_g) = M(\hat{C}_y) - \frac{K\,C_y^2\left[A_1^2C_x^2 + B_1A_2^2 - 2A_1A_2\,C_{12}(x,x)\right]}{\left[B_1C_x^2 - C_{12}^2(x,x)\right]} \tag{3.9}$$

where

$$T^* = (C_y, \overline{X},\ \sigma_x^2)$$

It is observed from (3.5), (3.6) and (3.9) that the asymptotically optimum estiamtors (AOE's) in d_g would be better than all the estimators in D_1 and D_2.

For large populations, we get, ignoring the terms $0(N^{-1})$,

$$A = 1,\ B = \frac{n-3}{n-1},\ C = \frac{2}{n-1},\ A - B - C = 0.$$

Since n is assumed to be large and terms $0(n^{-2})$ are neglected, in practice one may, in the above expressions, take A = 1, B = 1 and neglect the term C. In such a case, expressions (3.8) and (3.9) reduce to:

$$\left.\begin{aligned} g_2(T^*) &= \frac{C_y}{\overline{X}}\,\frac{\left[\sqrt{\beta_1(x)}A_1^*C_x - A_2^*\beta_2^*(x)\right]}{\left[\beta_2(x) - \beta_1(x) - 1\right]C_x^2} = \lambda_1^*\,\frac{C_y}{\overline{X}} \quad \text{(say)} \\ g_3(T^*) &= \frac{C_y\left[A_2^*\sqrt{\beta_1(x)} - C_xA_1^*\right]}{\sigma_x^2C_x\left[\beta_2(x) - \beta_1(x) - 1\right]} = \lambda_2^*\,\frac{C_y}{\sigma_x^2} \ \text{(say)} \end{aligned}\right] \tag{3.10}$$

and

$$M_0(d_g) = M^*(\hat{C}_y) - \frac{\left[A_1^{*2}C_x^2 + \beta_2^*(x)A_2^{*2} - 2A_1^* - 2A_1^*A_2^*\sqrt{\beta_1(x)}\,C_x\right]}{\left[\beta_2(x) - \beta_1(x) - 1\right]}, \tag{3.11}$$

where $\beta_1(x) = \mu_3^2(x)/\mu_2^3(x)$, $\qquad A_1^* = \left[\frac{1}{2}\{C_{22}(y,x) - 1\} - C_{12}^*(y,x)\right]$

$$A_2^* = \left[\frac{1}{2}\left\{C_{21}^*(y,x) - \rho\, C_y C_x\right\}\right], \quad C_{21}^*(y,x) = \mu_{21}(y,x)/(\overline{X}\,\sigma_y^2),$$

$$C_{12}^*(y,x) = \frac{\mu_{12}(y,x)}{\overline{Y}\sigma_x^2}, \qquad \lambda_1^* = \frac{\left[\sqrt{\beta_1(x)}\, A_1^* C_x - A_2^*\beta_2^*(x)\right]}{C_x^2\left\{\beta_2(x) - \beta_1(x) - 1\right\}},$$

$$\lambda_2^* = \frac{\left[A_2^*\sqrt{\beta_1(x)}\, A_1^* C_x\right]}{C_x^2\left\{\beta_2(x) - \beta_1(x) - 1\right\}}, \qquad \beta_1(y) = \frac{\mu_3^2(y)}{\mu_2^3(y)},$$

$$M^*(\hat{C}_y) = \frac{C_y^2}{n}\left[C_y^2 - \sqrt{\beta_1(y)}\, C_y + \frac{1}{4}\left\{\beta_2(y) - 1\right\}\right].$$

It is to be noted that $g_2(T^*) = \lambda_1$ and $g_3(T^*) = \lambda_2$ are rarely known in practice, hence it is advisable to replace these by their estimated optimum values based on the sample observations. We may take $\hat{g}_2(T^*)$ and $\hat{g}_3(T^*)$ as estimators of $g_2(T^*)$ and $g_3(T^*)$ where

$$\hat{g}_2(T^*) = \frac{\overline{X}\hat{C}_y}{\sigma_x^2}\frac{\left(\frac{m_3(x)}{\overline{X}\sigma_x^2}\right)\hat{A}_1^* - A_2^*\left(\frac{m_4(x)}{\sigma_x^4} - 1\right)}{\hat{\Delta}}$$

$$= \hat{\lambda}_1^*\frac{\hat{C}_y}{\overline{X}} = \hat{\lambda}_1 \text{ (say)} \tag{3.12a}$$

$$\hat{g}_3(T^*) = \frac{\overline{X}\hat{C}_y}{\sigma_x^3}\frac{\left(\frac{m_3(x)}{\sigma_x^3}\right)\hat{A}_2^* - A_1^*\left(\frac{\sigma_x}{\overline{X}}\right)}{\hat{\Delta}}$$

$$= \hat{\lambda}_2^*\frac{\hat{C}_y}{\sigma_x^2} = \hat{\lambda}_2 \text{ (say)} \tag{3.12b}$$

where

$$\hat{\Delta} = \left[\frac{m_4(x)}{\sigma_x^4} - \frac{m_3^2(x)}{\sigma_x^6} - 1\right],$$

$$\hat{\lambda}_1^* = \frac{\overline{X}^2}{\sigma_x^2} \frac{\left[\frac{m_3(x)}{\sigma_x^3} \hat{A}_1^* - \hat{A}_2^* \left\{ \frac{m_4(x)}{\sigma_x^4} - 1 \right\} \right]}{\hat{\Delta}},$$

$$\hat{\lambda}_2^* = \frac{\overline{X}}{\sigma_x} \frac{\left[\frac{m_3(x)}{\overline{X}\sigma_x^2} \hat{A}_2^* - \hat{A}_1^* \frac{\sigma_x}{\overline{X}} \right]}{\hat{\Delta}},$$

$$\hat{A}_1^* = \left[\frac{1}{2} \left\{ \frac{m_{22}(y,x)}{s_y^2 \, \sigma_x^2} - 1 \right\} - \frac{m_{12}(y,x)}{\overline{y} \, \sigma_x^2} \right],$$

$$\hat{A}_2^* = \left[\frac{1}{2} \frac{m_{21}(y,x)}{s_y^2 \overline{X}} - \frac{m_{11}(y,x)}{\overline{y} \, \overline{X}} \right]$$

where $m_{rs}(y,x) = \sum_{i=1}^{n} (y_i - \overline{y})^r (x_i - \overline{x})^s / n, \quad (r, s = 1,2,3,4)$

Now, we find the MSE in case of estimated optimum values λ_1^* and λ_2^* as follows :

From (3.10) and the regularity conditions for d_g, we want a function $g(\hat{C}_y, \overline{x}, s_x^2, \overline{X}, \sigma_x^2)$ such that

$$g^*(T^*) = C_y, \qquad g_1(T^*) = \left. \frac{\partial g}{\partial \hat{C}_y} \right|_{T^*} = 1,$$

$$g_2(T^*) = \left. \frac{\partial g}{\partial \overline{x}} \right|_{T^*} = \lambda_1, \; g_3(T^*) = \left. \frac{\partial g}{\partial s_x^2} \right|_{T^*} = \lambda_2 .$$

This means the function g will involve not only $(\hat{C}_y, \overline{x}, s_x^2, \overline{X}, \sigma_x^2)$ but λ_1 and λ_2 as well, and thus we want a function $g^*(\hat{C}_y, \overline{x}, s_x^2, \overline{X}, \sigma_x^2, \lambda_1, \lambda_2)$ such that

$$g^*(T^{**}) = C_y, \qquad g_1(T^{**}) = \left. \frac{\partial g^*}{\partial \hat{C}_y} \right|_{T^{**}} = 1,$$

$$g_2(T^{**}) = \left.\frac{\partial g^*}{\partial \bar{x}}\right|_{T^{**}} = \lambda_1, \quad g_3(T^{**}) = \left.\frac{\partial g^*}{\partial s_x^2}\right|_{T^{**}} = \lambda_2 .$$

Since in such a function $g^*(\hat{C}_y, \bar{x}, s_x^2, \bar{X}, \sigma_x^2, \lambda_1, \lambda_2)$ so required, λ_1 and λ_2 are unknown, we may take

$$g^{**}\hat{C}_y, \bar{x}, s_x^2, \bar{X}, \sigma_x^2, \hat{\lambda}_1, \hat{\lambda}_2) = g^*(\hat{C}_y, \bar{x}, s_x^2, \bar{X}, \sigma_x^2, \hat{\lambda}_1, \hat{\lambda}_2) .$$

Now

$$g^{**}(T^{**}) = C_y, \qquad g_1^{**}(T^{**}) = \left.\frac{\partial g^{**}}{\partial \hat{C}_y}\right|_{T^{**}} = 1,$$

$$g_2^{**}(T^{**}) = \left.\frac{\partial g^{**}}{\partial \bar{x}}\right|_{T^{**}} = \lambda_1, \text{ and } \quad g_3^{**}(T^{**}) = \left.\frac{\partial g^{**}}{\partial s_x^2}\right|_{T^{**}} = \lambda_2 .$$

We may take $d_g^* = g^{**}(\hat{C}_y, \bar{x}, s_x^2, \bar{X}, \sigma_x^2, \hat{\lambda}_1, \hat{\lambda}_2)$ as an estimator for C_y.

Expanding $g^{**}(\hat{C}_y, \bar{x}, s_x^2, \bar{X}, \sigma_x^2, \hat{\lambda}_1, \hat{\lambda}_2)$ about the point $T^{**}(C_y, \bar{X}, \sigma_x^2, \lambda_1, \lambda_2)$, we have

$$\begin{aligned} d_g^* &= g^{**}(T^{**}) + (\hat{C}_y - C_y)\, g_1^{**}(T^{**}) + (\bar{x} - \bar{X}) g_2^{**}(T^{**}) \\ &\quad + (s_x^2 - \sigma_x^2)\, g_3^{**}(T^{**}) + (\hat{\lambda}_1 - \lambda_1) \left.\frac{\partial g^{**}}{\partial \hat{\lambda}_1}\right|_{T^{**}} \\ &\quad + (\hat{\lambda}_2 - \hat{\lambda}_2) \left.\frac{\partial g^{**}}{\partial \hat{\lambda}_2}\right|_{T^{**}} + ... \\ &= \hat{C}_y + (\bar{x} - \bar{X})\lambda_1 + (s_x^2 - \sigma_x^2)\lambda_2 + (\hat{\lambda}_1 - \lambda_1)\, g_4^{**}(T^{**}) \\ &\quad + (\hat{\lambda}_2 - \lambda_2) g_5^{**}(T^{**}) + ... \end{aligned}$$

where $g_4^{**}(T^{**}) = \left.\frac{\partial g^{**}}{\partial \hat{\lambda}_1}\right|_{T^{**}}$ and $g_5^{**}(T^{**}) = \left.\frac{\partial g^{**}}{\partial \hat{\lambda}_2}\right|_{T^{**}}$

the first partial derivatives of g**(.) with respect to $\hat{\lambda}_1$ and $\hat{\lambda}_2$ at T** respectively are equal to zero so that MSE of d_g^* to order of n^{-1} becomes equal to minimum MSE $M_0(d_g)$ given in (3.11). Thus, if we take the estimator

$$d_g^* = g^{**}(\hat{C}_y, \bar{x}, s_x^2, \hat{\lambda}_1, \hat{\lambda}_2, \overline{X}, \sigma_x^2) \tag{3.13}$$

depending on estimated optimum values such that

$$\left.\begin{aligned} & g^{**}(C_y, \overline{X}, \sigma_x^2, \lambda_1, \lambda_2) = C_y \\ & g_1^{**}(C_y, \overline{X}, \sigma_x^2, \lambda_1, \lambda_2) = 1 \\ & g_2^{**}(C_y, \overline{X}, \sigma_x^2, \lambda_1, \lambda_2) = \lambda_1 \\ & g_3^{**}(C_y, \overline{X}, \sigma_x^2, \lambda_1, \lambda_2) = \lambda_2 \\ & g_4^{**}(C_y, \overline{X}, \sigma_x^2, \lambda_1, \lambda_2) = 0 \\ \text{and} \quad & g_5^{**}(C_y, \overline{X}, \sigma_x^2, \lambda_1, \lambda_2) = 0 \end{aligned}\right\} \tag{3.14}$$

The estimator d_g^* attains the minimum MSE $M_0(d_g)$ given in (3.11).
It may easily be verified that some particular cases:

(i) $$d_g^{*(1)} = \hat{C}_y \left[\frac{\bar{x}}{\overline{X}}\right]^{\hat{\lambda}_1^*},$$

(ii) $$d_g^{*(2)} = \hat{C}_y \left[1 + \hat{\lambda}_1^*\left(\frac{\bar{x}}{\overline{X}} - 1\right) + \hat{\lambda}_2^*\left(\frac{s_x^2}{\sigma_x^2} - 1\right)\right],$$

(iii) $$d_g^{*(3)} = \hat{C}_y \left[1 - \hat{\lambda}_1^*\left(\frac{\bar{x}}{\overline{X}} - 1\right) - \hat{\lambda}_2^*\left(\frac{s_x^2}{\sigma_x^2} - 1\right)\right]^{-1},$$

(iv) $$d_g^{*(4)} = \hat{C}_y + \hat{\lambda}_1(\bar{x} - \overline{X}) + \hat{\lambda}_2(s_x^2 - \sigma_x^2),$$

etc. attain minimum MSE $M_0(d_g)$ given in (3.11), where

$$\lambda_1^* = \frac{\overline{X}}{\sigma_x^2}\frac{\left[\hat{A}_1^* \frac{m_3(x)}{\overline{X}\sigma_x^2} - \hat{A}_2^*\left\{\frac{m_4(x)}{\sigma_x^4} - 1\right\}\right]}{\hat{\Delta}},$$

$$\lambda_2^* = \frac{\overline{X}}{\sigma_x^2}\frac{\left[\hat{A}_1^* \frac{m_3(x)}{\overline{X}\sigma_x^2} - \hat{A}_2^*\left\{\frac{m_4(x)}{\sigma_x^4} - 1\right\}\right]}{\hat{\Delta}},$$

$$\hat{\lambda}_2 = \hat{\lambda}_1^* \frac{\hat{C}_y}{\overline{X}} \text{ and } \hat{\lambda}_2 = \hat{\lambda}_2^* \frac{\hat{C}_y}{\sigma_x^2}.$$

4. Comparison of some estimators

In this section, we shall compare some particular members of the class of estimators d_g with the conventional estimators $\hat{C}_y$ of C_y.

We consider the following difference-type estimators of C_y :

$$d = \hat{C}_y - \delta_1(\bar{x} - \overline{X}) - \delta_2(s_x^2 - \sigma_x^2) \tag{4.1}$$

δ_i (i = 1, 2,) being suitably chosen constants.

For large N and n , ignoring finite population correction (fpc) term, the MSE of d to the terms of order $0(n^{-1})$ is given by

$$M(d) = M^*(\hat{C}_y) + \frac{1}{n}\left[\delta_1^2\sigma_x^2 + \delta_2^2\,\sigma_x^4\beta_2^*(x) + 2\delta_1\delta_2\overline{X}\sigma_x^2\sqrt{\beta_1(x)}C_x - 2\delta_1\overline{X}C_y\,A_2^* - 2\delta_2\sigma_x^2 C_y A_1^*\right] \tag{4.2}$$ where

$$M^*(\hat{C}_y) = \frac{C_y^2}{n}\left[C_y^2 - \sqrt{\beta_1(y)}\,C_y + \frac{\beta_2^*(y)}{4}\right]$$

If δ_i is pre-assigned, then the estimator d would be better than

$$d^{(1)} = \hat{C}_y - \delta_1(\bar{x} - \overline{X}) \tag{4.3}$$

iff δ_2 lies between 0 and 2 δ_{20}, where

$$\delta_{20} = \frac{\left[C_y A_1^* - \delta_1 \sqrt{\beta_1(x)}\, \sigma_x\right]}{\beta_2^*(x)\sigma_x^2} \tag{4.4}$$

is the optimum value of δ_2 in this case.

Then, to our order of approximation

$$d_1^* = d^{(1)} - \hat{\delta}_{20}(s_x^2 - \sigma_x^2) \tag{4.5}$$

will be uniformly better than $d^{(1)}$ whatever be the specified δ_1, where

$$\hat{\delta}_{20} = \frac{\left[\hat{C}_y \hat{A}_1^* \sigma_x^2 - \delta_1 m_3(x)\right]}{\left\{m_4^*(x) - \sigma_x^4\right\}} \tag{4.6}$$

such that $E(\hat{\delta}_{20}) = \delta_{20} + 0(n^{-1})$.

If δ_2 is pre-assigned, then the estimator d would be more efficient than

$$d^{(2)} = \hat{C}_y - \delta_2(s_x^2 - \sigma_x^2) \tag{4.7}$$

iff d_1 lies between 0 and 2 δ_{10}, where

$$\delta_{10} = \frac{\overline{X} C_y A_2^* - \delta_2 \mu_3(x)}{\sigma_x^2} \tag{4.8}$$

is the optimum value of δ_1 in this case.

For large samples,

$$\hat{\delta}_{10} = \frac{\overline{X}\, \hat{C}_y \hat{A}_2^* - \delta_2 m_3(x)}{\sigma_x^2} \tag{4.9}$$

has δ_{10} as its mean, i.e. $E(\hat{\delta}_{10}) = \delta_{10} + 0\,(n^{-1})$, and

$$d_2^* = d^{(2)} - \hat{\delta}_{10}(\overline{x} - \overline{X}) \tag{4.10}$$

would always be better than $d^{(2)}$ for a given δ_2. The reduction in MSE is given by

$$M(d^{(2)}) - M(d_2^*) = \frac{1}{n} \frac{\left[C_y \overline{X} A_2^* - \delta_2 \mu_3(x)\right]}{\sigma_x^2} \tag{4.11}$$

Further the estimator $d^{(1)}$, $d^{(2)}$ and

$$d^{(3)} = \hat{C}_y - \delta_3(\hat{C}_x - C_x) \tag{4.12}$$

would be better than the usual estimator $\hat{C}_y$ iff

$$0 < \delta_1 < 2\,\delta_{10}^* \tag{4.13}$$

$$0 < \delta_2 < 2\,\delta_{20}^* \tag{4.14}$$

and

$$0 < \delta_3 < 2\,\delta_{30}^* \tag{4.15}$$

respectively, where

$$\delta_{10}^* = \overline{X}\,C_y A_2^*/\sigma_x^2,$$

$$\delta_{20}^* = C_y A_1^*/\left(\sigma_x^2 \beta_2^*(x)\right)$$

and

$$\delta_{30}^* = -\left[\frac{C_y}{C_x}\right](Q^*/2)$$

are the optimum values of δ_1, δ_2 and δ_3 in (4.3), (4.7) and (4.12) respectively, where

$$Q^* = \frac{2A_2^* - A_1^*}{h^*(x)}$$

and

$$h^*(x) = \frac{\left[\beta_2^*(x) + 4C_x(C_x - \sqrt{\beta_1(x)}\right]}{4}.$$

It can easily be shown, to the first degree of approximation, that the estimator

$$d = \hat{C}_y - \hat{\delta}_1^*(\overline{x} - \overline{X}) - \hat{\delta}_2^*(s_x^2 - \sigma_x^2) \tag{4.16}$$

would be better than the estimators

$$d^{*(1)} = \hat{C}_y - \hat{\delta}_{10}^*(\overline{x} - \overline{X})$$

$$d^{*(2)} = \hat{C}_y - \hat{\delta}_{20}^*(s_x^2 - \sigma_x^2)$$

and

$$d^{*(3)} = \hat{C}_y - \hat{\delta}^*_{30}(\hat{C}_x - C_x)$$

where

$$\hat{\delta}_1 = \frac{\overline{X}\hat{C}_y}{\hat{\Delta}\sigma_x^2}\left[\frac{m_3(x)}{\sigma_x^3}\hat{A}_2^* - \hat{A}_1^*\frac{\sigma_x}{\overline{X}}\right],$$

$$\hat{\delta}_2 = \frac{\overline{X}\hat{C}_y}{\hat{\Delta}\sigma_x^3}\left[\frac{m_3(x)}{\overline{X}\sigma_x^2}\hat{A}_1^* - \hat{A}_2^*\left\{\frac{m_4(x)}{\sigma_x^4} - 1\right\}\right],$$

$$\hat{\delta}^*_{10} = \frac{\overline{X}\hat{C}_y\hat{A}_2^*}{\sigma_x^2},$$

$$\hat{\delta}^*_{20} = \frac{C_y\hat{A}_1^*\sigma_x^2}{\left[m_4(x) - \sigma_x^4\right]},$$

and

$$\hat{\delta}^*_{30} = -\frac{\overline{X}\hat{C}_y}{\sigma_x}\frac{\hat{Q}^*}{2}.$$

Further, it can easily be shown, to the first degree of approximation, that the estimators $d^{*(1)}$, $d^{*(2)}$ and $d^{*(3)}$ would always be better than

$$\left(\hat{C}_y, \hat{C}_y\frac{\overline{X}}{x}, \hat{C}_y\frac{\overline{x}}{\overline{X}}\right), \left(\hat{C}_y, \hat{C}_y\frac{\sigma_x^2}{s_x^2}, \hat{C}_y\frac{s_x^2}{\sigma_x^2}\right) \text{ and } \left(\hat{C}_y, \hat{C}_y\frac{C_x}{\hat{C}_x}, \hat{C}_y\frac{\hat{C}_x}{C_x}\right)$$

respectively.

Remark 4.1: In case of a bivariate normal population the equation (3.10) and (3.11) reduce to:

$$g_2(T^*) = \rho\, C_y^2/\sigma_x \tag{4.17 a}$$

$$g_3(T^*) = \rho^2\, C_y/\sigma_x^2 \tag{4.17 b}$$

and

$$M_0(d_g) = M^{**}(\hat{C}_y) - \frac{1}{n}\rho^2 C_y^2\left(C_y^2 + \frac{1}{2}\rho^2\right) \tag{4.18}$$

where

$$M^{**}(\hat{C}_y) = \frac{1}{n} C_y^2 \left(\frac{1}{2} + C_y^2\right) \tag{4.19}$$

Further, the equations (3.5) and (3.6) reduce to:

$$M_0(D_1) = M^{**}(\hat{C}_y) - \frac{\rho^2 C_y^4}{n} \tag{4.20}$$

$$M_0(D_2) = M^{**}(\hat{C}_y) - \frac{\rho^2 C_y^4}{2n} \frac{(\rho + 2C_y C_x)^2}{(1 + 2C_x^2)} \tag{4.21}$$

It follows from (4.18), (4.20) and (4.21) that the estimator d_g is more efficient than D_1 and D_2 reported by Das and Tripathi (1981 a,b).

5. Empirical study

The following two populations are considered to illustrate the relative performance of various estimators of C_y.

Population I : It consists of 142 cities of India with population (number of persons) 1,00,000 and above; the characters x and y being

x: census population in the year 1961 (in 00's)

y : census population in the year 1971 (in 00's)

Values of the required population parameters are given below:

$\overline{Y} = 4015.2183$, $\overline{X} = 2900.3872$, $C_y = 2.1118$, $C_x = 2.1971$, $\rho = 0.9948$,

$C_{12}(y,x) = 13.0450$, $C_{21}(y,x) = 13.2294$, $C_{12}(y,y) = 12.4981$,

$C_{12}(x,x) = 14.0330$, $C_{22}(y,x) = 43.7615$, $\beta_2(y) = 40.8536$,

$\beta_2(x) = 48.1567$, N = 142, n = 30, A = 1.0139, B = 0.9593,

K = 0.0265, C = 0.0466.

Population II : Data under consideration were taken from census (1961), West Bengal, District Census Hand Book, Midnapore. The population consists of 353 villages or town/ward under Panskura Police Station. The characters y and x are number of persons, area of villages or towns/wards, in acres respectively. For this population required values of the parameters are:

$\overline{Y} = 670.0764$, $\overline{X} = 274.3728$, $C_y = 0.9586$, $C_x = 0.7338$, $\rho = 0.7528$,

$C_{12}(y,x) = 2.1905$, $C_{21}(y,x) = 1.6301$, $C_{12}(y,y) = 2.6861$,

$C_{12}(x,x) = 1.9937$, $C_{22}(y,x) = 12.3060$, $\beta_2(y) = 15.05$,

$\beta_2(x) = 16.3895$, N = 353, n = 30, A = 1.0054, B = 0.9422, K = 0.03059,

C = 0.0634

[*Source :* Das and Tripathi (1981 b)]

Table 1 gives the relative efficiencies of the various estimators of C_y with respect to conventional estimator $\hat{C}_y$

Table 1. Relative efficiencies (%) of various estimators of C_y with respect to $\hat{C}_y$

Estimators	**Population I**	**Population II**
$\hat{C}_y$	100.00	100.00
D_1	100.86	166.26
$D_1^{(1)} = \hat{C}_y(\bar{x}/\bar{X})$	61.60	18.95
$D_1^{(2)} = \hat{C}_y(\bar{X}/\bar{x})$	98.20	71.45
D_2	199.90	1900.46
$D_2^{(1)} = \hat{C}_y(C_x^2/\hat{C}_x^2)$	31.70	54.71
$D_2^{(2)} = \hat{C}_y(C_x/\hat{C}_x)$	140.28	1179.76
$D_2^{(3)} = \hat{C}_y(\hat{C}_x^2/C_x^2)$	10.25	9.25
$D_2^{(4)} = \hat{C}_y(\hat{C}_x/C_x)$	24.90	21.86
$D_2^{(5)} = \hat{C}_y - (\hat{C}_x^2/C_x^2)$	115.37	1.26
$D_2^{(6)} = \hat{C}_y$ - $(\hat{C}_x/C_x)$	186.82	964.76
$e_4^{(1)} = \hat{C}_y\left(\frac{\bar{x}}{\bar{X}}\right)\left(\frac{\hat{\sigma}_x^2}{\sigma_x^2}\right)^{\alpha}$	189.95 (α_{opt} = - 0.3542)	1443.25 (α_{opt} = - 0.47409)
$e_4^{(2)} = \hat{C}_y\left(\frac{\bar{x}}{\bar{X}}\right)\left(\frac{\hat{\sigma}_x^2}{\sigma_x^2}\right)^{\alpha}$	110.48 (α_{opt} = - 0.09367)	90.18 (α_{opt} = - 0.11225)
$e_4^{(3)} = \hat{C}_y\left(\frac{\bar{x}}{\bar{X}}\right)^{\alpha}\left(\frac{\hat{\sigma}_x^2}{\sigma_x^2}\right)$	12.12 (α_{opt} = - 4.23282)	14.82 (α_{opt} = - 3.32115)
$e_4^{(4)} = \hat{C}_y\left(\frac{\bar{x}}{\bar{X}}\right)^{\alpha}\left(\frac{\sigma_x^2}{\hat{\sigma}_x^2}\right)$	36.51 (α_{opt} = - 3.17234	79.16 (α_{opt} = - 2.49293)

Estimators	Population I	Population II
$e_4^{(5)} = \hat{C}_y \left(\frac{\bar{x}}{\bar{X}}\right)^{\alpha} \left(\frac{\hat{\sigma}_x}{\sigma_x}\right)$	29.08 (α_{opt} = - 2.38153)	35.41 (α_{opt} = - 1.86763)
$e_4^{(6)} = \hat{C}_y \left(\frac{\bar{x}}{\bar{X}}\right)^{\alpha} \left(\frac{\hat{\sigma}_x}{\sigma_x}\right)$	146.69 (α_{opt} = - 1.32105)	1232.24 (α_{opt} = 1.03941
$e_4^{(7)} = \hat{C}_y \left(\frac{\bar{x}}{\bar{X}}\right)\left(\frac{\sigma_x}{\hat{\sigma}_x}\right)$	149.62	1179.75
$e_4^{(8)} = \hat{C}_y \left(\frac{\bar{X}}{\bar{x}}\right) \left(\frac{\hat{\sigma}_x}{\sigma_x}\right)$ Das and Tripathi (1992)	42.62	21.86
$e_5 = \frac{\hat{\sigma}_y}{\left[\bar{y} - \hat{\beta}(\bar{x} - \bar{X})\right]}$ ($\hat{\beta}$ is the sample regression coefficient of y on x)	102.35	77.86
$e_6 = \frac{\left[\hat{\sigma}_y - \alpha(\hat{\sigma}_x - \sigma_x)\right]}{\bar{y}}$	146.69 (α_{opt} = 1.32105)	406.98 (α_{opt} = 1.03941)
$e_7 = \frac{\left[\hat{\sigma}_y - (\hat{\sigma}_x - \sigma_x)\right]}{\bar{y}}$	167.74	89.23
d_g	201.85	2040.92

Table 1 demonstrates that the efficiency of d_g is considerably high (more) than that of others in both the populations.

REFERENCES

DAS, A. K. (1988): Contributions to the theory of sampling strategies based on auxiliary information. Ph. D. thesis submitted to BCKV, Mohanpur, Nadia, West Bengal, India.

DAS, A. K. and TRIPATHI, T. P. (1980): Sampling strategies for population mean when conefficient of variation of an auxiliary character is known. Sankhya, 42, C, 76-86.

DAS, A. K. and TRIPATHI, T. P. (1977): Admissible estimators for quadratic forms in finite populations. Bull. Inter. Stat. Inst., 47, Book 4, 132-135.

DAS, A. K. and TRIPATHI, T. P. (1978): Use of auxiliary information in estimating the finite population variance, Sankhya, C, 40, 139-148.

DAS, A. K. and TRIPATHI, T. P. (1981 a): Sampling strategies for coefficient of variation using knowledge on the mean of an auxiliary character. Tech. Rep. Stat. Math. 5/81, ISI, Calcutta.

DAS, A. K. and TRIPATHI, T. P. (1981 b): A class of estimators for coefficient of variation using knowledge on coefficient of variation of an auxiliary character. Paper presented at 35th annual Conference of Ind. Soc. Agricultural Statist. Held at New Delhi, India.

DAS, A. K. and TRIPATHI, T. P. (1981 c): A class of sampling strategies for population mean using information on mean and variance of an auxiliary character. Proceedings of Indian Statistical Institute, Golden Jubilee International Conference on Statistics: Applications and new directions, 174-181.

DAS, A. K. and TRIPATHI, T. P. (1992-93): Use of auxiliary information in estimating the coefficient of variation. Alig. J. of Statist., 12 and 13, 51-58.

LIU, T. P. (1974): A general unbiased estimator for the variance of finite population. Sankhya, C, 36, 23-32.

SINGH, H. P., UPADHYAYA, L. N. and NOMJOSHI, U. D. (1988): Estimation of finite population variance. Current Science, 57, 24, 1331-1334.

SINGH, H.P., UPADHYAYA, L. N. and IACHAN, R. (1990): An efficient class of estimators using supplementary information in sample surveys. Alig. J. of Statist., 10, 37-50.

SINGH, J., PANDEY, B. N. and HIRANO, K. (1973): On the utilization of a known coefficient of kurtosis in the estimation procedure of variance. Ann. Inst. Statist. Math., 25, 51-55.

SEARLS, D. T., and INTERAPANICH, P. (1990): A note on an estimator for the variance that utilizes the kurtosis. The American Statistician, 44(40), 195-296.

SRIVASTAVA, S. K. (1971): A generalized estimator for the mean of a finite population using multi auxiliary information. J. Amer. Statist, Assoc., 66, 404-407.

SRIVASTAVA, S. K. (1980): A class of estimators using auxiliary information in sample surveys. Canad. J. Statist., 8(2), 253-254.

SRIVASTAVA, S. K. and JHAJJ, H. S. (1986): On the estimation of finite population correlation coefficient. J. Ind. Soc. Agricultural. Statist., 38(10), 82-91.

SRIVASTAVA, S. K. and JHAJJ, H. S. (1981): A class of estimators of the population mean in survey sampling using auxiliary information. Biometrika, 68(1), 341-343.

UPADHYAYA, L. N. and SINGH, H. P. (1985): A class of estimators using auxiliary information for estimating ratio of two finite population means. Guj. Statist. Rev., 12, 2, 7-16.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 1103–1104

Book Review

W. Charemza and K. Strzala (Eds):
East European Transition and EU Enlargement:
A Quantitative Approach, Physica-Verlag
Heidelberg, New York, 2002, 384p.

The major aim of this collection of papers selected by Charemza and Strzala is to show that a variety of considerations – both theoretical and empirical- enter into the process of analysis of the possible accession of some transition economies of Eastern Europe in the present European Union. The collection is based on a conference organised by the editors. They are highly eminent, distinguished, with very good track records as social scientists, who have done a very good job in getting some of the quality papers together in the Gdansk conference of June, 2001.

I am inclined to think that it is a very timely and important contribution to the growing literature on the growth and development of transition economies. Most of the papers chosen by the editors are of high quality. Quite a few of them display evidence of high quality and thorough research. Further, many articles have useful policy implications. The story lines from different countries are inevitably variable and so are the policy implications. Yet the book allows readers to detect the commonalities and contrasts among country-specific cases. There is a clear emphasis on research in empirical time series economics and the use of relevant econometric techniques of rigorous order. Hence, as a research in topical, significant and relevant Eastern and Western European economic issues, I strongly support the publication of this book by Springer mainly because I feel that it has achieved its main objective. Although there are a few 'rival' books in this particular area, I still have no hesitation in saying that some of the papers are of very good quality and could easily have been published in good economic journals.

I don't think that the reviewed publication would be a text book, but, it would be a valuable and excellent source of reference to students, particularly graduate students, researchers, teachers and administrators. I would certainly

recommend it to my students in European Economics and Economies of Transition. As a secondary text, the book has a niche market that could be exploited fully.

Subrata Ghatak
Research Professor in Economics
Kingston University, Surrey, KT1 2EE, UK

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp.1105–1116

ECONOMIC AND SOCIAL TRENDS IN TRANSITION COUNTRIES

BASIC ECONOMIC TRENDS IN COUNTRIES OF CENTRAL AND EASTERN EUROPE AND CIS COUNTRIES

1. Gross Domestic Product

Following the economic slowdown in 2001, in the I half of 2002 a further weakening in the rate of annual GDP growth was noted in several countries of Central and Eastern Europe. This involved Poland, Slovenia and Hungary. In the case of Poland, following a considerable weakening of the dynamic of GDP in the II half of 2001, in the I half of 2002 a small acceleration in the rate of GDP growth was noted, while in the other two countries, a further slowing was noted. However, Bulgaria, Slovakia, Estonia and Lithuania noted a higher dynamic of GDP in 2002 (particularly in the II quarter) than in the previous year. An improvement in the economic dynamic has been observed in the Czech Republic and Romania since 2001, following a decline in GDP in 1997-1999 and a small growth in 2000. In the I half of 2002, GDP growth in these countries remained relatively high, but in the Czech Republic somewhat lower and in Romania - similar to that noted in 2001. Among countries of the discussed group, Poland noted the smallest annual GDP growth in the I half of 2002, while it was somewhat higher in the II quarter than in the I quarter (0.8%, against 0.5%). In the remaining countries, GDP growth in the II quarter of 2002 ranged from 2.5% in the Czech Republic to 7.0% in Estonia. Moreover, in the majority of countries this growth was higher than in the I quarter.

In CIS countries as a whole, following a significant acceleration in the dynamic of GDP in 2000, the rate of GDP growth slowed in 2001, with this trend continuing in the I half of 2002. Among others, this was due to the weakening in GDP growth in Russia, from 9.0% in 2000 to 5.0% in 2001 and 3.9% in the I half of 2002. However, Armenia, Kazakhstan, Tajikistan, Turkmenistan and Ukraine have noted a successive acceleration in the dynamic of GDP in recent years, while in the I half of 2002 the dynamic of annual GDP weakened somewhat in these countries (excluding Armenia) in comparison with that noted in 2001. Armenia, Azerbaijan, Kazakhstan and Tajikistan noted the highest GDP growth in the I half of 2002 (in the 8-10% range), while Georgia and Russia noted the lowest growth (in the 3-4% range). In the I half of 2002, a decline in GDP was observed in Kyrgyzstan.

Gross domestic product (constant prices)

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= =100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
CEFTA									
Bulgaria	89.8	93.0	103.5	102.3	105.4	104.0	96.9	103.2	105.3
Czech Republic	104.9	99.0	98.8	99.6	100.5	103.3	110.1	102.8	102.5
Hungary	101.4	104.6	104.9	104.2	105.2	103.8	126.6	102.9	103.1
Poland	106.0	106.8	104.8	104.1	104.0	101.0	129.7	100.5	100.8
Slovenia	103.5	104.6	103.8	105.2	104.6	103.0	127.3	102.2	103.2
Romania	104.0	93.9	95.2	98.8	101.8	105.3	98.6	103.1	105.7
Slovak Republic	106.5	106.2	104.1	101.3	102.2	103.3	125.8	103.9	104.0
Baltic countries									
Estonia	103.9	110.6	104.7	99.4	107.1	105.0	134.5	103.2	107.0
Latvia	103.3	108.6	103.9	102.8	106.8	107.7	138.0	103.8	104.9
Lithuania	104.7	107.3	105.1	96.1	103.8	105.9	124.6	104.4	106.9
CIS[a)]	**96.8**	**101.0**	**96.4**	**104.6**	**108.3**	**106.0**	**113.2**	**..**	**104.0**
Armenia	105.9	103.3	107.3	103.3	106.0	109.6	140.9	107.4	110.1
Azerbaijan	101.3	105.8	110.0	107.4	111.1	109.9	154.6	104.7	108.4
Belarus	102.8	111.4	108.4	103.4	105.8	104.1	141.3	104.0	104.7
Georgia	111.2	110.6	102.9	103.0	102.0	104.5	139.0	103.7	103.7
Kazakhstan	100.5	101.7	98.1	102.7	109.8	113.2	128.0	110.4	109.2
Kyrgyzstan	107.1	109.9	102.1	103.7	105.4	105.3	138.3	97.2	95.1
Moldova	94.1	101.6	93.5	96.6	102.1	106.1	93.6	104.8	106.4
Russian Federation	96.6	100.9	95.1	105.4	109.0	105.0	111.8	103.7	103.9
Tajikistan	83.3	101.7	105.3	103.7	108.3	110.2	110.4	109.3	108.3
Turkmenistan	106.7	88.6	107.1	116.9	117.6	120.5	167.6	..	..
Ukraine	90.0	97.0	98.1	99.8	105.9	109.1	98.6	104.1	104.4
Uzbekistan	101.7	105.2	104.4	104.4	103.8	104.5	126.4	103.1	104.2

a) Quarters 2002 – increasingly.

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, Economic Survey of Europe, 2002/1, unece, Geneva, Statistical Bulletin 10/2002, CSO, Warsaw, www.cisstat.com.

In the 1996-2001 period, among countries of Central and Eastern Europe, the highest growth in GDP was noted in Latvia (almost 40%), Estonia (approximately 35%) and Poland (approximately 30%). In Poland this was due to the high dynamic of GDP in the initial years of the discussed period, while in the other two countries, a high dynamic also in the final years of the period. Bulgaria and Romania noted a drop in GDP in 2001 in relation to 1995. In CIS countries as a whole, GDP growth in the discussed period was substantially lower than in the majority of countries of Central and Eastern Europe, amounting to approximately 13%. This was due to a decline in GDP in 1996 and 1998 as well as a small growth in 1997, in turn due to, among others, poor economic results during this period in Russia, Moldova and Ukraine. Among CIS countries, the highest GDP growth in the 1996-2001 period (in the range of approximately 40-70%) was noted in Armenia, Azerbaijan, Belarus, Georgia and Turkmenistan, the lowest (in the range of approximately 10-25%) was noted in Russia, Tajikistan and Uzbekistan. Moldova and Ukraine noted a decline in GDP in 2001 in relation to 1995.

2. Industrial Output

In the majority of countries of Central and Eastern Europe, in the I half of 2002, a weakening in the annual dynamic of industrial output was observed in comparison with that noted the previous year. Poland was the only country of the discussed group in which the level of industrial output in the I half of 2002 was lower than the previous year, following a small decline also observed in 2001. In the remaining countries of this group in the II quarter of 2002, an acceleration was noted in the growth rate of output, the highest in the Baltic countries: Lithuania (up to approximately 8%), Estonia (almost 7%) and Latvia (up to approximately 6%). Slovenia and the Czech Republic also noted a relatively high growth in industrial output in the II quarter of 2002 (above 5% and approximately 5%, respectively).

Industrial output (constant prices)

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= =100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
CEFTA									
Bulgaria	105.1	90.1	92.1	90.3	110.3	97.6	84.7	91.0	103.2
Czech Republic	102.0	104.5	101.6	96.9	105.4	106.8	118.1	104.1	104.9
Hungary	103.4	111.1	112.5	110.4	118.1	103.6	174.6	100.6	101.6
Poland	108.3	111.5	103.5	103.6	106.7	99.9	138.1	98.4	99.6
Slovenia	101.0	101.0	103.7	99.5	106.2	102.9	115.0	101.7	102.5
Romania	106.3	92.8	86.2	97.6	107.1	108.2	96.2	103.1	104.0
Slovak Republic	102.5	101.3	105.4	97.3	108.6	106.9	123.6	101.1	105.4
Baltic countries									
Estonia	103.5	115.2	103.2	96.6	114.5	107.8	146.6	98.4	106.8
Latvia	105.4	113.7	103.2	94.6	104.7	106.9	130.8	100.1	105.8
Lithuania	104.1	104.5	108.3	88.8	105.3	116.9	128.7	101.7	108.1
CIS[a)]	**96.0**	**103.0**	**98.0**	**108.0**	**111.0**	**107.0**	**124.3**	**..**	**104.0**
Armenia	101.0	101.0	98.0	105.0	105.9	104.0	115.6	113.9	112.1
Azerbaijan	93.0	100.3	102.0	104.0	107.0	105.0	111.2	100.1	101.5
Belarus	104.0	119.0	112.0	110.0	108.0	106.0	174.7	102.0	104.0
Georgia	107.0	108.0	98.0	107.0	111.0	95.0	127.8	101.2	100.1
Kazakhstan	100.3	104.0	98.0	103.0	116.0	114.0	139.2	112.1	108.7
Kyrgyzstan	104.0	140.0	105.0	96.0	106.0	105.0	163.4	88.1	86.4
Moldova	94.0	100.0	85.0	88.0	108.0	114.0	86.5	108.8	111.0
Russian Federation	96.0	102.0	95.0	111.0	112.0	105.0	121.4	102.6	103.2
Tajikistan	76.0	98.0	108.0	106.0	110.0	115.0	107.9	105.4	107.9
Turkmenistan	120.0	77.9	102.1	115.1	130.0	111.0	158.6	..	..
Ukraine	95.0	99.7	99.0	104.0	112.0	114.0	124.6	103.1	105.8
Uzbekistan	102.6	104.1	103.6	105.7	105.9	107.6	133.2	..	..

a) Quarters 2002 – increasingly.

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, Economic Survey of Europe, 2002/1, unece, Geneva, Statistical Bulletin 10/2002, CSO, Warsaw, www.cisstat.com.

In CIS countries as a whole, following a high growth in industrial output in 2000 (11.0%) and slower growth in 2001 (7.0%), a further weakening in the growth rate (to 4.0%) was noted in the I half of 2002. This was primarily due to the results obtained in these years in countries with significant industrial potential (i.e., Russia, Belarus, Georgia and Ukraine). In all these countries (excluding Georgia) the annual dynamic of industrial output in the I half of 2002 was lower than in the previous year, with the greatest weakening noted in Ukraine. In the I half of 2002, the highest growth in industrial output (above 10%) was noted in Armenia and Moldova. In Russia, Belarus and Ukraine, the level of output was higher than in the I half of 2001, in the range of approximately 3-6%. However, a deep decline in output occurred in Kyrgyzstan, after noting relatively high growth in 2000 and 2001.

In relation to 1995, among countries of Central and Eastern Europe, Hungary, Estonia and Poland were characterised by the highest dynamic of industrial output in 2001 (a growth of more than 70%, more than 40% and almost 40%, respectively). In 2001, only Romania noted a lower level of output than in 1995, resulting from its steep drop in 1997-1999. However, among CIS countries in the 1996-2001 period, the largest growth in output was noted in Belarus (almost 75%), Kyrgyzstan (more than 60%) and Turkmenistan (almost 60%). In Russia and Ukraine the growth in output during the discussed period was significantly smaller (more than 20%), which was due to poor results obtained in industry during the initial years of this period.

3. Gross Agricultural Output

A significant growth in annual gross agricultural output was observed in 2001 in the majority of CEFTA group countries, following a decline in 2000. A high growth in output, on a level of approximately 23%, was noted in Romania (after an approximate 15% drop in 2000) and Hungary (approximate 14% growth), following a decline in 1999-2000. In the remaining countries of the discussed group (excluding Bulgaria) a growth in gross agricultural output (in the range of 2.5-8%) was also observed, with Slovakia noting the first growth since 1996.

Among Baltic countries, characterised by a severe drop in gross agricultural output in 1998-1999 and growth in 2000, in 2001 only Latvia noted higher agricultural output than the previous year (5.0%). Estonia and Lithuania noted a decline in 2001-9.7% and 5.4%, respectively.

In CIS countries as a whole, 2001 was the second consecutive year in which a relatively high growth in gross agricultural output was noted (in the range of 6-8%). Among others, this was due to good production results in Russia, Ukraine

and Belarus as well as in several Asian countries. A growth in agricultural output in 2001, higher than the previous year, was noted in all countries of the discussed group, while in Kazakhstan, Armenia, Azerbaijan, Tajikistan and Ukraine the dynamic of output was higher than in CIS countries as a whole.

Gross agricultural output (constant prices)

Country	1996	1997	1998	1999	2000	2001	
	previous year=100						1995= =100
CEFTA							
Bulgaria	82.4	106.9	98.5	102.7	90.9	99.7	80.8
Czech Republic	98.6	94.9	100.7	100.6	95.5	102.5	92.9
Hungary	106.3	96.2	102.9	98.9	94.1	114.3	112.0
Poland	100.7	99.8	105.9	94.8	94.4	105.7	100.6
Slovenia	100.7	98.8	102.2	98.7	102.4	..	102.8[a)]
Romania	101.3	103.4	92.5	104.0	85.2	122.7	105.3
Slovak Republic	102.0	99.0	94.1	97.5	87.7	107.8	87.5
Baltic countries							
Estonia	93.7	98.5	96.4	89.6	108.2	90.3	77.8
Latvia	94.3	102.0	92.1	89.4	104.7	105.0	87.0
Lithuania	112.6	108.6	94.8	85.5	105.4	94.6	98.9
CIS	**95.0**	**101.0**	**90.0**	**102.0**	**106.0**	**108.0**	**100.9**
Armenia	102.0	94.0	113.0	101.0	97.5	112.0	119.6
Azerbaijan	103.0	94.0	106.0	107.0	112.0	111.0	136.5
Belarus	102.0	95.0	99.3	92.0	109.0	102.0	98.4
Georgia	106.0	107.0	90.0	108.0	85.0	106.0	99.4
Kazakhstan	95.0	99.2	81.0	128.0	96.0	117.0	109.7
Kyrgyzstan	115.0	112.0	103.0	108.0	103.0	107.0	157.9
Moldova	87.0	112.0	88.0	92.0	97.0	104.0	79.5
Russian Federation	95.0	102.0	87.0	104.0	108.0	107.0	101.3
Tajikistan	91.0	100.2	106.0	103.0	113.0	111.0	124.9
Turkmenistan	87.0	..	..	..	..	..	..
Ukraine	91.0	98.0	90.0	93.0	110.0	110.0	90.4
Uzbekistan	94.0	106.0	104.0	106.0	103.0	105.0	118.8

a) 2000.

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, www.cisstat.com.

In the 1996-2001 period, among countries of all discussed groups, Kyrgyzstan noted the highest growth in agricultural output (approximately 60%). Among CEFTA group countries the largest growth in agricultural output in the discussed period was noted in Hungary (12%) and Romania (more than 5%), while Bulgaria noted the largest decline (approximately 20%). A decline in gross agricultural output in the period of the last six years also occurred in all Baltic countries (the deepest in Estonia). In CIS countries as a whole, despite a relatively high dynamic observed in the majority of countries in this group in 2000 and 2001, the growth in agricultural output in the 1996-2001 period amounted to only 0.9%. Among others, this was due to the drop in output in Moldova, Ukraine, Belarus and Georgia as well as a small growth in Russia.

4. Inflation

The annual growth in prices of consumer goods and services in the majority of countries of Central and Eastern Europe observed in successive quarters of 2002, was lower than that noted the previous year. Consumer prices in Bulgaria, Lithuania and Latvia grew somewhat faster in the I quarter of 2002 than in 2001, but in the II quarter inflation in these countries weakened considerably. Romania continued to note the highest growth in consumer prices in 2002 (above 20%), following high growth also noted in previous years. In the remaining countries of the discussed group inflation in 2001 and the I half of 2002 did not exceed 10% and in the II quarter of 2002 ranged from 0.5% in Lithuania to 7.6% in Slovenia. In the latter years of the 1996-2001 period, the growth rate of consumer prices in Lithuania and Latvia slowed significantly, remaining on a level below 3%. Among CEFTA group countries the lowest level of inflation in 2001 was noted in the Czech Republic (below 5%), and the highest (excluding Romania) in Hungary (almost 10%), while in successive years of the 1996-2001 period inflation in Hungary declined successively and this trend also continued in the I half of 2002.

Consumer price indices

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= 100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
CEFTA									
Bulgaria	221.7	1158.3	118.7	102.6	110.3	107.4	about 37 times	108.2	107.1
Czech Republic	108.9	108.4	110.6	102.1	103.9	104.7	144.9	103.7	102.3
Hungary	123.6	118.4	114.2	110.0	109.8	109.2	220.4	106.2	105.5
Poland	119.9	114.9	111.8	107.3	110.1	105.5	192.0	103.4	102.1
Slovenia	109.9	108.4	108.1	106.1	108.9	108.4	161.3	108.0	107.6
Romania	138.8	254.9	159.3	145.8	145.7	134.5	about 16 times	126.9	124.3
Slovak Republic	106.1	106.1	106.7	110.6	112.0	107.3	159.6	104.7	103.1
Baltic countries									
Estonia	123.1	111.1	110.6	103.3	104.0	105.8	172.0	104.3	104.2
Latvia	117.7	108.5	104.7	102.4	102.6	102.5	144.0	103.3	101.9
Lithuania	124.7	108.8	105.1	100.8	101.0	101.3	147.0	102.5	100.5
CIS [a)]									
Armenia	119.0	114.0	109.0	101.0	99.2	103.0	152.6	100.4	101.7
Azerbaijan	120.0	104.0	99.0	91.0	102.0	102.0	117.1	101.6	102.3
Belarus	153.0	164.0	173.0	394.0	269.0	161.0	about 74 times	147.0	145.6
Georgia	139.0	107.0	104.0	119.0	104.0	105.0	201.0	105.3	105.9
Kazakhstan	139.0	117.0	107.0	108.0	113.0	108.0	229.3	105.6	105.5
Kyrgyzstan	132.0	123.0	110.0	136.0	119.0	107.0	309.3	102.7	101.6
Moldova	124.0	112.0	108.0	139.0	131.0	110.0	300.4	106.0	106.0
Russian Federation	147.9	114.7	127.8	185.7	120.8	121.6	591.1	118.0	115.8

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= 100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
Tajikistan	370.0	172.0	143.0	126.0	124.0	137.0	about 19 times	106.4	107.0
Turkmenistan	814.0	183.7	116.8	..	..	..	..	..	..
Ukraine	180.0	116.0	111.0	123.0	128.0	112.0	408.7	103.7	102.2
Uzbekistan	154.0	158.8	117.7	129.0	124.9	..	463.9[b)]	..	..

a) Quarters 2002 – increasingly (except Russian Federation). b) 2000 .

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, Economic Survey of Europe, 2002/1, unece, Geneva, Statistical Bulletin 10/2002, CSO, Warsaw, OLIS-Net – database of OECD, www.cisstat.com.

In relation to 1995, in 2001 the highest growth in consumer prices was noted in Bulgaria (approximately 37-times, mainly as a result of their very high, almost 12-fold growth in 1997) and Romania (approximately 16-times, as a result of a significant growth, in the range of approximately 35-155%, in consecutive years of the discussed period). Consumer prices in the Czech Republic, Lithuania and Latvia noted the smallest growth in the 1996-2001 period (below 50%).

In many CIS countries in consecutive, particularly the final years of the discussed period, inflation was substantially higher than in the majority of countries of Central and Eastern Europe. A comparable growth in consumer prices was noted in Georgia as well as in several Asian countries. Very high, double- and triple-digit inflation has continued for several years in Belarus and Tajikistan, and has continued on a somewhat lower level (but much higher than in the Baltic countries in particular and in many countries of the CEFTA group) in Russia and Ukraine. In the I half of 2002 inflation declined in all of the mentioned CIS countries; to the greatest degree in Tajikistan and Ukraine. As a result, in 2002 the slowest growth in consumer prices was noted in Armenia, Azerbaijan, Kyrgyzstan and Ukraine (below 2.5%), and the fastest growth was noted in Belarus and Russia (approximately 46% and approximately 16%, respectively). In relation to 1995, in 2001 the largest growth in consumer prices was noted in Belarus (approximately 74-times) and Tajikistan (approximately 19-times), and the smallest growth was noted in Azerbaijan (about 17%). In the remaining CIS countries, the growth in consumer prices ranged from approximately 50% in Armenia to about 500% in Russia.

5. Unemployment

Unemployment in the majority of countries of Central and Eastern Europe in successive years of the 1996-2001 period was higher than in CIS countries (with the exception of Russia and Armenia), in which a low, single-digit rate of unemployment was observed in this period. Among CEFTA group countries, only

in Hungary did the unemployment rate systematically decline in the successive years of the discussed period, to 5.7% in 2001 and 5.6% in the II quarter of 2002 (i.e., to the lowest level among countries of this group). Countries in which a significant growth in unemployment was noted in recent years include Bulgaria, Poland and Slovakia. The unemployment rate in these countries in the II quarter of 2002 ranged from 17.5% in Bulgaria to 19.9% in Poland. The unemployment rate in the remaining countries of the CEFTA group during all years of the 1996 – 2001 period did not exceed 10% and in the II quarter of 2002 a further decline was noted (with the exception of Romania).

Unemployment rate

Country	1996	1997	1998	1999	2000	2001	2002	
							I Q.	II Q.
	in %							
CEFTA[a]								
Bulgaria	12.5	14.0	12.2	13.7	18.1	17.5	17.8	17.5
Czech Republic	3.9	4.8	6.5	8.7	8.8	8.1	7.7	7.0
Hungary	9.9	8.7	7.8	7.0	6.4	5.7	5.8	5.6
Poland	12.3	11.2	10.6	13.9[b]	16.1	18.2	20.3	19.9
Slovenia	7.3	7.4	7.8	7.6	7.0	6.4	6.9	5.9
Romania	6.8	6.0	6.3	6.8	7.1	6.6	10.0	8.0
Slovak Republic	11.1	11.6	11.9	16.2	18.6	19.2	19.4	18.6
Baltic countries[a]								
Estonia	9.9	9.6	9.8	12.2	13.6	12.6	11.2	9.4
Latvia	20.3	15.2	14.2	14.3	14.4	13.1	13.7	13.3
Lithuania	16.4	14.1	13.3	14.1	15.4	17.0	17.1	13.0
CIS[c]	**6.6**	**7.6**	**9.0**	**8.3**	**7.0**	**6.2**	**..**	**..**
Armenia	9.7	11.0	8.9	11.5	10.9	9.8	..	..
Azerbaijan	1.1	1.3	1.4	1.2	1.2	1.3	..	..
Belarus	4.0	2.8	2.3	2.0	2.1	2.3	2.6	2.6
Georgia	3.2	8.0	4.2	5.6	..	..	..	..
Kazakhstan	4.1	3.9	3.7	3.9	3.7	2.8	..	..
Kyrgyzstan	4.5	3.1	3.1	3.0	3.1	3.1	..	..
Moldova	1.5	1.7	1.9	2.1	1.8	1.7	..	..
Russian Federation[a]	10.0	11.2	13.3	12.2	9.8	9.0	..	..
Tajikistan	2.4	2.8	2.9	3.1	3.0	2.6	..	..
Turkmenistan	..	..	..	..	..	..	..	..
Ukraine	1.5	2.8	4.3	4.3	4.2	3.7	..	..
Uzbekistan	0.3	0.3	0.4	0.5	0.6	0.4	..	..

a) Unemployment rate by LFS. b) Calculated as average at I and IV quarter as in remaining quarters the survey was not realised. c) Registered unemployment rate.

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, Economic Survey of Europe, 2002/1, unece, Geneva, Statistical Bulletin 6/2002, Ministry of Statistics and Analysis of the Republic of Belarus, Minsk, Labour Force Survey in Poland 4/2001, CSO, Warsaw, Statistical Bulletin 10/2002, CSO, Warsaw, OLIS-Net – database of OECD.

Among Baltic countries the highest level of unemployment in 2001 was observed in Lithuania (17.0%) and the lowest in Estonia (12.6%). In the II quarter of 2002, the unemployment rate in these two countries declined significantly, to 9.4% in Estonia and to 13.0% in Lithuania, while it grew somewhat, to 13.3%, in Latvia.

In the successive years of the 1996-2001 period, the unemployment rate in CIS countries as a whole, did not exceed 9% and, following growth in the 1996-1998 period, in subsequent years fell to 6.2% in 2001. Among this group of countries the highest level of unemployment in 2001 was observed in Armenia (almost 10%) and Russia (9.0%), while it has successively declined in Russia since 1999. The lowest level of unemployment in all years of the discussed period was noted in Uzbekistan (below 1%), while in the remaining countries in 2001 it ranged from 1.3% in Azerbaijan to 3.7% in Ukraine.

6. Foreign Trade

A surplus of imports over exports was observed in successive years of the 1996-2001 period in the majority of CEFTA group countries, Baltic countries and CIS countries. This trend continued in the I half of 2002. Only in Russia and Kazakhstan, in all years of the discussed period, as well as in Azerbaijan, Turkmenistan, Uzbekistan and Ukraine, in the latter years, did exports exceed imports; approximately 2-fold in Russia in 1999-2002. Among CEFTA group countries the largest surplus of imports over exports was noted in the latter years in Bulgaria, Poland and Romania, among Baltic countries – in Latvia, and among CIS countries – in Armenia and Georgia. In the I half of 2002 the value of imports exceeded the value of exports to the smallest degree in the Czech Republic, Slovenia and Belarus.

Imports as a percent of exports

Country	1996	1997	1998	1999	2000	2001	2002	
							I Q.	II Q.
CEFTA								
Bulgaria	103.8	99.8	118.2	137.7	134.9	142.0	130.8	146.7
Czech Republic	126.5	120.5	109.8	107.0	110.7	109.4	102.9	104.7
Hungary	115.5	111.2	111.7	112.0	114.2	110.4	108.0	108.0
Poland	152.0	164.3	166.7	167.5	154.6	139.3	131.8	139.7
Slovenia	113.4	111.9	111.6	118.0	115.8	109.7	106.3	104.8
Romania	141.4	133.8	142.6	124.4	125.9	136.6	126.0	130.3
Slovak Republic	126.0	120.6	120.7	110.7	107.4	116.8	113.5	113.8
Baltic countries								
Estonia	155.4	151.4	147.9	143.8	133.8	129.8	137.3	141.4
Latvia	160.8	162.8	176.0	170.9	170.7	175.1	165.8	175.4
Lithuania	135.9	146.2	156.1	161.0	143.3	138.6	139.9	146.8
CIS								
Armenia	295.2	382.8	408.1	349.6	294.0	256.4	170.2	192.6

Country	1996	1997	1998	1999	2000	2001	2002	
							I Q.	II Q.
Azerbaijan	152.3	101.7	177.7	111.5	67.2	61.8	78.2	77.0
Belarus	122.8	119.0	120.9	112.9	118.0	108.8	106.8	101.3
Georgia	345.2	393.3	458.0	252.9	197.3	213.8	279.7	209.5
Kazakhstan	71.7	66.2	80.0	65.9	55.3	73.6	71.5	81.2
Kyrgyzstan	165.9	117.4	163.8	132.2	109.7	98.1	122.7	106.5
Moldova	134.8	133.9	162.0	126.5	164.4	157.4	153.2	160.9
Russian Federation	78.6	83.6	78.4	52.2	42.5	52.9	56.4	56.6
Tajikistan	86.8	100.5	119.1	96.2	86.1	105.5	113.0	99.2
Turkmenistan	60.1	157.5	169.7	126.1	71.2	83.3	..	..
Ukraine	122.2	120.3	116.1	102.3	95.8	97.0	91.6	96.4
Uzbekistan	111.9	104.0	97.1	93.8	88.2	97.1	..	..

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, Economic Survey of Europe, 2002/1, unece, Geneva, International Financial Statistics, 2002/7, IMF, Washington D.C., Statistical Bulletin 10/2002, CSO, Warsaw, www.cisstat.com.

In many countries of Central and Eastern Europe, following a decline in 1999, since 2000 a growth has been observed in both annual exports and imports. However, in 2001 this growth was smaller than the previous year. Among countries of the discussed group Hungary was characterised by the steadiest growth in turnover in individual years of the 1996–2001 period. However, in 2001 exports grew the fastest (to a greater degree than imports) in Lithuania, the Czech Republic and Poland, while imports grew the fastest (significantly more than exports) in Romania. In the I quarter of 2002 a decline in annual foreign trade turnover, particularly in regard to imports, was noted in many countries of Central and Eastern Europe, while in the II quarter turnover again increased. In CEFTA group countries the growth in exports was higher than the growth in imports. The most rapid growth rate in exports was observed in Romania and the Czech Republic, while Lithuania and Latvia noted the most rapid rate of growth in regard to imports.

Indices of exports (in USD)

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= =100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
CEFTA									
Bulgaria	91.5	101.0	84.9	95.5	120.4	106.0	95.6	93.6	101.6
Czech Republic	103.0	104.0	113.7	101.3	110.5	115.2	157.1	102.1	115.8
Hungary	122.0	121.6	120.4	108.7	112.3	108.6	236.9	102.7	110.9
Poland	106.7	105.4	109.6	97.1	115.5	114.0	157.7	98.5	112.6
Slovenia	99.9	100.7	108.1	94.4	102.2	106.0	111.2	96.6	111.8
Romania	102.2	104.3	98.5	102.2	122.2	109.8	143.9	100.9	117.5
Slovak Republic	102.8	109.3	111.8	94.9	115.8	106.6	147.2	104.6	108.3
Baltic countries									
Estonia	113.1	141.1	110.3	73.6	132.9	104.6	180.0	76.5	96.2

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= =100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
Latvia	110.7	115.9	108.3	95.1	108.4	107.1	153.4	99.2	108.1
Lithuania	124.0	115.1	96.1	90.9	126.8	120.3	190.1	101.0	109.1
CIS									
Armenia	107.0	80.3	94.8	105.0	129.7	113.6	126.0	156.6	133.2
Azerbaijan	99.1	123.8	77.6	153.3	187.8	132.6	363.3	61.0	100.4
Belarus	117.7	129.2	96.8	83.6	124.0	102.2	156.0	96.3	107.1
Georgia	130.9	120.6	80.4	123.3	138.7	97.7	210.8	..	..
Kazakhstan	112.6	109.9	83.7	102.9	163.2	94.8	164.8	94.3	87.8
Kyrgyzstan	123.5	119.6	85.1	88.3	111.2	94.3	116.4	98.7	114.3
Moldova	106.6	110.1	72.2	73.4	101.7	120.8	76.5	103.4	98.9
Russian Federation	109.7	100.8	83.7	102.5	139.5	96.2	127.4	87.0	99.4
Tajikistan	102.8	96.9	80.0	115.4	113.8	83.2	87.1	..	..
Turkmenistan	89.4	44.6	79.1	200.3	210.1	108.0	143.6	..	..
Ukraine	109.7	98.8	88.8	91.7	125.8	111.6	124.0	101.0	101.9
Uzbekistan	149.3	95.6	79.9	99.4	100.9	96.3	110.1	..	..

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, Economic Survey of Europe, 2002/1, unece, Geneva, International Financial Statistics, 2002/7, IMF, Washington D.C., Statistical Bulletin 10/2002, CSO, Warsaw, www.cisstat.com.

In the majority of CIS countries, following a significant growth in foreign trade turnover in 2000, in 2001 a weakening in this growth or a decline was observed, particularly in regard to exports. Only in Moldova, on the side of exports, as well as in Azerbaijan, Russia, Tajikistan, Turkmenistan and Uzbekistan, on the side of imports, was the dynamic of turnover in 2001 higher than the previous year. The largest decline in exports among countries of the discussed group occurred in 2001 in Tajikistan, while Kyrgyzstan noted the largest decline in imports. In the I half of 2002 a further decline in turnover was observed in many of these countries, among others, in regard to exports, in Russia (with a relatively high growth in imports) and Kazakhstan (also with a decline in imports). In 2002, the largest growth in exports was noted in Armenia, while Kyrgyzstan noted the largest growth in imports.

Indices of imports (in USD)

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= =100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
CEFTA									
Bulgaria	90.0	97.2	100.5	111.3	112.0	111.6	122.2	96.1	100.6
Czech Republic	109.7	99.1	103.5	98.7	114.4	113.6	144.3	97.1	112.7
Hungary	117.3	117.0	121.1	109.0	114.5	105.0	217.6	98.9	108.0
Poland	127.8	113.9	111.2	97.6	106.6	102.7	172.9	95.9	109.4
Slovenia	99.3	99.4	107.8	99.9	100.3	100.3	106.9	95.0	104.4
Romania	111.3	98.6	104.9	89.2	123.7	119.1	151.3	98.9	106.9
Slovak Republic	126.6	104.6	111.9	87.0	112.4	116.0	168.1	95.4	107.8

Country	1996	1997	1998	1999	2000	2001		2002	
	previous year =100						1995= =100	I Q. corresp. period of previous year=100	II Q. corresp. period of previous year=100
Baltic countries									
Estonia	127.2	137.4	107.8	71.6	123.6	101.5	169.2	84.4	108.3
Latvia	127.6	117.4	117.1	92.4	108.3	109.8	192.8	103.3	116.2
Lithuania	125.0	123.8	102.7	83.4	112.9	116.4	174.3	105.6	120.9
CIS									
Armenia	127.0	104.2	101.1	89.9	109.1	99.1	130.0	114.4	93.6
Azerbaijan	144.1	82.6	135.6	96.2	113.1	122.1	214.4	109.3	110.1
Belarus	124.7	125.2	98.4	78.1	129.6	94.2	146.5	101.5	100.5
Georgia	186.2	137.4	93.6	68.1	108.1	105.1	185.2	..	..
Kazakhstan	111.4	101.4	101.1	84.8	137.0	126.0	167.1	99.9	94.5
Kyrgyzstan	160.5	84.6	118.8	71.3	92.3	84.3	89.4	137.1	126.5
Moldova	127.5	109.3	87.4	57.3	132.2	115.6	106.7	108.5	102.2
Russian Federation	112.9	107.2	78.5	68.3	113.5	119.8	88.3	109.4	108.5
Tajikistan	82.5	112.3	94.8	93.2	101.8	101.9	84.9	..	..
Turkmenistan	74.1	117.0	85.2	148.8	118.7	126.4	165.1	..	..
Ukraine	113.7	97.3	85.7	80.7	117.8	113.0	101.8	98.8	107.2
Uzbekistan	171.5	88.8	74.7	96.0	95.0	106.0	109.9	..	..

Source: CESTAT Statistical Bulletin 2001/4, CSO, Warsaw, CANSTAT Statistical Bulletin 2002/2, CSO, Warsaw, Economic Survey of Europe, 2002/1, unece, Geneva, International Financial Statistics, 2002/7, IMF, Washington D.C., Statistical Bulletin 10/2002, CSO, Warsaw, www.cisstat.com.

In the 1996–2001 period, among countries of all discussed groups, Hungary, Azerbaijan and Georgia were characterised by the highest growth in exports (ranging from approximately 110% to 260%), while the dynamic of exports in these countries significantly exceeded the dynamic of imports. In 2001, a decline in the level of exports, in relation to the level noted in 1995, occurred in Bulgaria, Moldova and Tajikistan, while Kyrgyzstan, Russia (with an approximate 30% growth in exports) and Tajikistan noted a decline in the level of imports.

M. Bieńkowska, E. Czumaj, J. Gniadzik

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 1117–1119

Obituary

Tore E. Dalenius
1917 - 2002

The statistical profession lost one of its prominent members when Professor Tore E. Dalenius, passed away at the age of 84.

Tore Dalenius was born in 1917 in Jukkasjärvi in Northern Sweden and studied at the universities of Stockholm and Uppsala. His doctoral thesis, "*Sampling in Sweden*," was published in 1957. In Sweden, it was at the time considered a breakthrough for sample surveys based on a solid theoretical framework. His professional career included a number of positions and affiliations with agencies and institutions such as Cornell University, Bell Labs, IBM, the U.S. Department of Commerce, Statistics Sweden, and the U.S. Bureau of the Census. He ended his career as a Visiting Professor at Brown University in Providence, Rhode Island, U.S.A.

One of Prof. Dalenius's greatest achievements was the solution of the stratification problem, which was published in a series of articles in *Skandinavisk Aktuarietidsskrift* and *JASA* between 1952 and 1959 (some of them written with J.L. Hodges, Jr.). This was the big issue for sampling experts at the time when Neyman's and Tschuprow's solution to the allocation problem was well known. Prof. Dalenius showed how to determine the number of strata and also where to put the stratum boundaries. Applying although Prof. Dalenius was fully aware of the potential of sampling theory, he was also well aware of its limitations. As expressed by Prof. Dalenius and other prominent survey statisticians of the era, including P.C. Mahalanobis, W.E. Deming, and M.H. Hansen, sampling was a theory for only a restricted number of parameters. It was a theory for finite populations, not considering the processes that generate them, and it was a theory for true values, not including measurement errors. The concerns about those limitations led to the creation of the International Association of Survey Statisticians (IASS). IASS first appeared at the Vienna session of the International Statistical Institute in 1973, where Prof. Dalenius organized an invited paper session on how to balance different sources of errors in the design of surveys, an important issue that still needs a lot of attention. First time I met Prof. Dalenius at the Vienna Session in 1973. He included my paper entitled "*Sampling and Non-sampling Errors in Poland*" to the invited paper session mentioned above. From

that meeting we met several times at different ISI sessions discussing some aspects of sample surveys in Poland focusing on non-sampling errors.

As a Professor at Stockholm University, with an obligation to focus on official statistics problems, Prof. Dalenius initiated and ran two major research projects under the auspices of the Bank of Sweden Tercentenary Foundation. Both projects were extremely productive. The first project, "*Errors in Surveys*," focused on different sources of error in surveys and made fruitful attempts at gathering those errors under one survey model. Prof. Dalenius had a unique sense for interdisciplinary cooperation in research. Research was performed in fields such as medical diagnosis and experimental psychology to find statistical models that could be used in survey work. In the late 1960s, he had already cooperated with the Institute of Psychology at the University of Stockholm to describe the cognitive processes that generate respondents' answers to interview and self-administered questionnaires and how these processes affect response quality. He even taught classes on this subject. Thus, he was many years ahead of his contemporaries. Today most distinguished statistical agencies have a cognitive laboratory. Protecting the integrity of survey respondents was another major interest of Prof. Dalenius, and this became the topic of his second major research project. Again he was far ahead of most of his contemporaries. He reformulated the basic effectiveness criterion for a good survey design, "*To minimize the mean squared error at a given cost*," by adding "*and with maintained integrity for those who provide the data*." Today, the law in most countries protects respondents' integrity. Prof. Dalenius had a strong understanding of the importance of international relations for the exchange of ideas and new research results. He was a frequent visitor to other countries, and during the early 1980s he was chair of the American Statistical Association's committee for international relations in statistics. He worked particularly hard to increase the opportunities for survey statisticians in developing countries to take part in international conferences and to have access to survey literature. He was a consultant and advisor to organizations such as the United Nations and the World Health Organization. In the early 1970s, he moved permanently to the U.S. when he married Marjory Ann Schmink in Providence, Rhode Island. Prof. Dalenius was very active in enforcing the role of statistics in society and appeared frequently in the media, in Sweden and abroad. He was a straightforward critic of bad survey practice, and, in particular, he led a debate about the sloppy acceptance of large nonresponse rates in surveys.

He also published a number of well-written popular science articles in different fields of statistics such as Monte-Carlo Sampling and Queuing Theory.

Having Prof. Dalenius as a teacher and mentor meant hard work but was extremely rewarding. While he was a Professor at Stockholm University, there was a constant flow of prominent survey statisticians visiting the Department of Statistics. His bird's eye view of the current status of survey methodology was enormous. He was a devoted survey statistician, normally working no less than 70

hours a week during his active professional life. He always delivered on time; many of his documents were finished on weekends and holidays. He was an excellent correspondent, always providing well considered responses and interesting viewpoints. Prof. Dalenius spent his last decades in Rhode Island, gradually slowing down his activities in the company of his wife and visiting old friends and colleagues. Although he suffered from health problems, it seems as if the last decade of his life was peaceful and enjoyable. Looking at the current status of survey methodology, it is amazing how much it has been influenced by Prof. Dalenius. His main viewpoint, which he conveyed to his colleagues and students, was the importance of good data for society's decision making. We are proud and happy to have worked with one of the great minds during an era of very important survey developments.

One of the founders of the IASS Tore Dalenius was the Association's first Scientific Secretary from 1973 to 1977, the Chairman of the Program Committee for the 1975 session, and President from 1981 to 1983.

Prepared on the basis the Survey Statistician, vol. 46, July 2002, pp. 4-5, by Jan Kordos.

STATISTICS IN TRANSITION, December 2002
Vol. 5, No. 6, pp. 1121–1123

ACKNOWLEDGEMENTS

Referees of Volume 5

The Editorial Board wishes to thank the following referees who have generously given their time and skills to the *Statistics in Transition* during the period from January 2001 to December 2002, i.e. for preparing Volume 5.

Nibia Aires, Sahlgrenska, Gothenburg University, Sweden

Czesław Bracha, Warsaw School of Economics, Warsaw, Poland.

Rolf Dahlin, Mid Sweden University, Sweden

Stig Danielsson, Linköping University, Sweden

Czesław Domański, University of Łódź, Poland

Sławomir Dorosiewicz, Warsaw School of Economics, Warsaw, Poland

Erling Englund, R&D-Centre Sundsvall, Sweden

Ewa Frątczak, Central Statistical Office of Poland, Warsaw, Poland.

Wayne A. Fuller, Iowa State University, USA

Marek Fuchs, Eichsstaett, Germany

Elżbieta Gołata, University of Economics, Poznań, Poland.

Marek Góra, Warsaw School of Economics, Warsaw, Poland

Marek Gruszczyński, Warsaw School of Economics, Warsaw, Poland.

Johan Heldal, Statistics Norway, Norway

Anders Holmberg, Statistics Sweden, Sweden

Krzysztof Jajuga, Wrocław University of Economics, Wrocław, Poland

Sven-Erik Johansson, Karolinska Medical University, Sweden

Graham Kalton, WESTAT, Inc. Washington, USA

Irena Kasperowicz-Ruka, Warsaw School of Economics, Warsaw, Poland.

Visi Kivisniemi, Jyväskylä University, Finland
Jan Kordos, Warsaw School of Economics, Warsaw, Poland.
Irena Kotowska, Warsaw School of Economics, Warsaw, Poland
Liliana Kursa, Central Statistical Office, Warsaw, Poland
Seppo Laaksonen, Statistics Finland
Janis Lapins, Bank of Latvia, Riga, Latvia
Risto Lehtonen, Jyväskylä University, Finland
Hans Malker, R&D-centre Sundsvall, Sweden
Małgorzata Misztal, , University of Łódź, Łódź, Poland
Marek Męczarski, Warsaw School of Economics, Warsaw, Poland.
Wojciech Niemiro, Warsaw University, Warsaw, Poland
Lennart Nordberg, Statistics Sweden, Sweden
Lucyna Nowak, Central Statistical Office of Poland, Warsaw, Poland.
Jerzy Nowakowski, Warsaw School of Economics, Warsaw, Poland.
Paul Ollila, Statistics Finland, Finland
Tomasz Panek, Warsaw School of Economics, Warsaw, Poland.
Jan Paradysz, University of Economics, Poznań, Poland
Dariusz Parys, , University of Łódź, Łódź, Poland
Dorota Pekasiewicz, University of Łódź, Łódź, Poland
Richard Platek, Formerly Statistics Canada, Ottawa, Canada
Jarosław Podgórski, Warsaw School of Economics, Warsaw, Poland.
Waldemar Popiński, Central Statistical Office, Warsaw, Poland
Krystyna Pruska, University of Łódź, Łódź, Poland
J.N. K. Rao, Dept.of Maths and Statistics, Carleton University, Canada
Martin Ribe, Statistics Sweden, Sweden
Teresa Słaby, Warsaw School of Economics, Warsaw, Poland.
Honorata Sosnowska, Warsaw School of Economics, Warsaw, Poland
Adam Szulc, Warsaw School of Economics, Warsaw, Poland
Mirosław Szreder, University of Gdańsk, Gdańsk, Poland
Daniel Thorburn, Stockholm University, Sweden